統計数理(2012)
第 60 巻 第 2 号 289–303

特集「多様性と進化の統計解析」
［研究ノート］

配列組成の不均一性が分子系統解析の頑健性に及ぼす影響：タンパク質コード遺伝子を想定したシミュレーションによる評価

石川 奏太[†]・橋本 哲男[†]

(受付 2012 年 1 月 11 日；改訂 6 月 13 日；採択 6 月 15 日)

要　　旨

遺伝子配列の塩基・アミノ酸組成が生物間において極端に不均一であるとき，組成の均一性を仮定する一般的な分子系統解析は誤った進化系統樹の推測結果をもたらすことが多い．本研究ではこの問題を定量的に評価するために，タンパク質コード遺伝子に基づく解析を想定し，シミュレーション実験による検証を行った．まず，コドンの第一，第二，第三座位それぞれのアデニン＋チミン含量(AT 含量)が生物間で不均一なタンパク質コード遺伝子の配列データおよびモデル系統樹を用意し，コドン置換モデルを用いて最尤法により枝長などのパラメータを推定した．その際，系統樹上で各コドン座位の塩基組成が不均一であることを許容する置換モデルを使用した．最尤推定されたパラメータを用いて，コドン置換モデルに基づきモンテカルロシミュレーション法による配列生成を行うことで，塩基組成およびアミノ酸組成が配列間で不均一な仮想的配列データを生成した．得られたデータに対し，各系統での配列組成の均一性を仮定する一般的な置換モデルによる解析を行った結果，塩基配列・アミノ酸配列いずれに基づく解析でも，真の系統樹は殆ど復元されず，AT 含量やアミノ酸組成の類似した配列が互いに近縁関係にあることを示す系統樹(アーティファクト)が高頻度で誘導された．本研究は，タンパク質コード遺伝子における生物間での配列組成の不均一性が一般的な分子系統解析の精度にどれほどの悪影響を与えるかという問題に対し，初めての定量的評価を与えたものである．

キーワード：分子系統樹の推測，最尤法，塩基・アミノ酸組成の不均一性，コドン置換モデル，シミュレーション，モデル不整合．

1. はじめに

進化生物学や系統分類学では，生物の進化の歴史を推測するために，塩基(DNA・RNA)配列やアミノ酸配列，コドン配列などの遺伝子配列から進化系統樹を再構築する「分子系統解析」と呼ばれる手法が広く用いられている．近年，網羅的発現遺伝子解析(EST 解析)などのゲノム解析技術の飛躍的な進歩に伴い，あらゆる生物の遺伝子配列データの取得が容易になったことで，高次の分類群を超え，広範囲にわたる生物群を対象とした大域的な分子系統解析を行うことが可能となった．しかしながらその一方で，多様な生物種由来の遺伝子配列を分子系統解析に用いる場合，従来の解析方法では頑健な推測結果を得られないような事例が多く存在するこ

[†] 筑波大学 生命環境科学研究科：〒305–8572 茨城県つくば市天王台 1–1–1

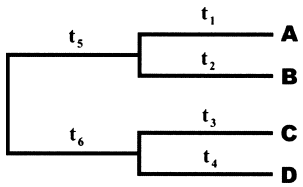

図 1. 4 配列からなる系統樹．A，B，C，D は各末端配列，t は系統樹の各枝を示す．

とも明らかとなってきた．こうした問題は，主として一般的な分子系統解析方法で仮定される「配列進化プロセスの系統間での均一性」と，実配列における進化プロセスとの不整合によって生じる．本稿ではとくに，塩基組成やアミノ酸組成など，配列の組成値の生物間での不均一性が，均一を仮定する従来の解析方法の頑健性に与える影響について，シミュレーション実験の結果を踏まえて方法論的観点から論じたい．

現状最も一般的に用いられている最尤法による分子系統解析では，遺伝子配列の進化は各系統で独立なマルコフ過程に従うことが仮定される．すなわち，ある塩基やアミノ酸，コドンなどの形質状態が別の形質状態に遷移する確率は現在の状態のみに依存し，どのようにして現在の状態に至ったかという過去には依存しないというものである．最尤法による分子系統解析では現存する遺伝子配列を材料とし，マルコフ過程を仮定し，その遷移確率から系統樹の尤度を計算・比較することで最大尤度をもつ系統樹(最尤系統樹)を推測する．ここで，簡単のため図 1 に示したような 4 種の遺伝子配列からなる系統樹の尤度計算について説明する．

n 個の座位からなる遺伝子配列 A，B，C，D の h 番目の座位において形質状態 p，q，r，s が観察され，配列 A，B および配列 C，D の共通祖先における形質状態をそれぞれ i，j とする．また，各形質状態から次の形質状態に遷移する時間は系統樹の各枝 $t_1 \sim t_6$ で表現されるとする．すると，ある座位 h における図 1 の系統樹の尤度 $L(\theta|X_h)$ は次のように表現される．

$$L(\theta|X_h) = f(X_h|\theta) = \sum_i \pi_i p_{ip}(t_1) p_{iq}(t_2) \sum_j p_{ij}(t_5+t_6) p_{jr}(t_3) p_{js}(t_4) \tag{1.1}$$

ここで，π_i は A，B，C，D の実配列データより推定される形質状態 i の出現頻度である．$p_{ij}(t)$ は時間 t における i から j への遷移確率を示す(後述するが，この遷移確率は i から j への瞬間置換速度を定義した置換モデル Q によって計算される)．h 番目の座位における尤度は，祖先配列の形質状態 i,j の組み合わせ全てについての総和を求めることで算出される．なお，遺伝子配列の各座位の置換は他の座位とは独立に起こると仮定することにより，最終的な系統樹の尤度 $L(\theta|X)$ は各座位における尤度を総乗することで求められる．

$$L(\theta|X) = \prod_{h=1}^{n} L(\theta|X) \tag{1.2}$$

ここで注目すべきことは，遷移確率 P を計算するために用いられる置換モデル Q は，系統樹上の全ての枝で同一のものが適用されるという点である(これを均一な置換モデルと呼ぶ)．そのため一般的な分子系統解析では，配列の進化は全ての系統で独立しているにも関わらず，その置換プロセスは均一な置換モデルによって表現されることが前提となっている．従って，上に記した尤度もまた「系統樹上で全ての配列が同じ置換モデルに従って進化する」という前提より計算されるものである．このように系統樹上の配列進化をただ一つの置換モデルによって表現する利点は，一つには尤度計算に必要なパラメータの推定にかける計算時間を短縮できるという点である．しかしながら，多様な生物種由来の配列を解析に用いた際，いずれかの配列

が他とは異なる置換プロセスに従って進化した場合には，均一な置換モデルを使用した分子系統解析ではモデルの不整合により真の系統樹とは異なる系統樹（アーティファクト）が誘導される危険性が生じる．

特に，塩基・アミノ酸・コドン配列のいずれにも関わらず，殆どの置換モデル（Q）において，形質状態 i から j への瞬間置換速度を表現するために置換後の形質状態の出現頻度 π_j が使用される．そのため，均一な置換モデルを尤度計算に使用した場合，系統樹上の全ての枝においてある形質状態 j の出現頻度は均一であり，配列の組成は各系統で一定に保たれる，という仮定が置かれることになる．しかしながらこの仮定は，実在する遺伝子配列を使った分子系統解析では非現実的なものとなる場合が多い．例えば，真正細菌のゲノムではアデニン＋チミン含量（AT 含量）が生物間で不均一であることが知られ，Mycoplasma 属細菌のゲノムでは 75%程度である一方，Micrococcus 属細菌では 25%程度である．また，これら真正細菌の 16S リボソーム RNA 配列における AT 含量も 20～50%と大小様々である（Mooers and Holmes, 2000）．異なる生物種間での AT 含量の不均一性は，後生動物のミトコンドリアゲノムにおいても確認されている．昆虫や線虫などのミトコンドリアゲノムにおける AT 含量は 65～85%程度であるのに対し，軟体動物では 60～71%，鳥類やほ乳類，硬骨魚類では 54～68%である（Saccone et al., 1999）．これらの例に加え，真核生物由来の核ゲノムにおいても同様の事例は確認されている（Chang and Campbell, 2000; Tarrio et al., 2001）．このことから，塩基組成，特に AT 含量の不均一性は実データ解析では極めて普遍的に確認される問題であることが分かる．

このように生物間で塩基組成に不均一性が認められる配列データを分子系統解析に使用する際には，全ての配列が同一の頻度パラメータ π に従って進化するという前提は適切ではないため，モデルの不整合によるアーティファクトが誘導される可能性が非常に大きい．事実，塩基組成が不均一であるリボソーム RNA 配列に均一な塩基置換モデルを適用した分子系統解析では，「進化的距離が大きく離れていても，塩基組成が類似している生物は近縁である」とみなす系統樹が誤推測された（Lockhart et al., 1994）．また，同様の問題はタンパク質コーディング遺伝子を用いた塩基置換モデルによる解析でも生じることが報告されている（Chang and Campbell, 2000）．

さらに，タンパク質コーディング遺伝子では各コドン座位における塩基組成の偏りはコドンの使用頻度の偏りとも密接な関係がある（Stenico et al., 1994）．従って，塩基組成に不均一性がある場合には，塩基配列翻訳後のアミノ酸配列の組成にも不均一性が生じ，均一なアミノ酸置換モデルを用いた分子系統解析に悪影響を及ぼすことが考えられる．その一例を図 2 に示す．

図 2 は真核生物の色素体およびその起源となったシアノバクテリアに由来する 7 遺伝子（*rpl14*, *rpl16*, *rps3*, *rps11*, *rps12*, *rpoB*, *tufA*）の配列データをもとに塩基置換モデルおよびアミノ酸置換モデルを用いて行った系統解析の結果である．これによると，アピコンプレクサ類は緑藻起源の二次葉緑体をもつユーグレナ藻類および *Helicosporidium* sp. と姉妹群を形成し，それは塩基配列に基づく解析およびアミノ酸配列に基づく解析双方において高いブートストラップ値で支持されている．このことから，アピコンプレクサ類がもつ退化型の色素体である「アピコプラスト」は緑藻由来のオルガネラであることが示唆される．しかしながら近年の研究で，アピコンプレクサ類の祖先的生物種であり光合成能のある *Chromera velia* が発見され，その色素体が紅藻起源であることが明確に示されたため，アピコプラストが紅藻由来のオルガネラであることはほぼ決定的となっている（Moore et al., 2008; Janouškovec et al., 2010）．すなわち，色素体配列を用いた系統解析ではアピコンプレクサ類は紅藻起源色素体をもつ生物（特に珪藻類や渦鞭毛藻類）と近縁となるべきであり，図 2 の系統樹は明らかに誤った系統樹だということになる．

ここで，図 2 の系統解析に用いたデータセットでは各コドン座位において塩基組成の不均

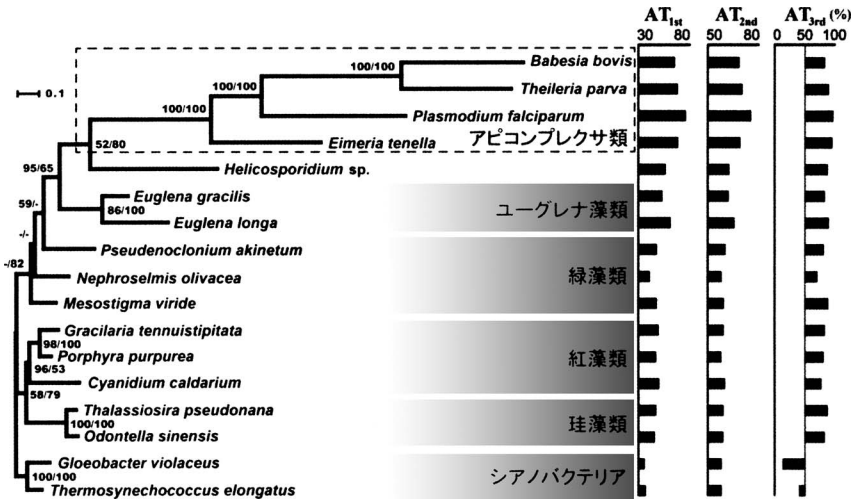

図 2. 色素体由来 7 遺伝子に基づく最尤法系統解析．樹形はアミノ酸配列に基づく解析より得られた最尤系統樹を示す．各ノードの値は左から①コドンの全座位を使用した塩基配列に基づく解析．②コドン配列を翻訳したアミノ酸配列に基づく解析それぞれのブートストラップ値(100 反復)を示す．系統解析には RAxML Ver.7.2.8(Stamatakis et al., 2008) を使用し，置換モデルは塩基・アミノ酸配列それぞれに GTR+Γ，LG+Γ を使用した．

一性がみられ，アピコンプレクサ類，ユーグレナ藻類，*Helicosporidium* sp. では，コドンの全座位において AT 含量が極端に高いことが確認できる．このように各コドン座位における塩基組成に極端に偏りのある遺伝子配列ではコドンの使用頻度にも偏りが生じるため，翻訳後のアミノ酸配列においてもアミノ酸組成に不均一性が生じることとなる．そのような配列では，特に AAA，AAT，ATT，TTT，TTA，TAA，TAT，ATA などのコドンの使用頻度が高くなり，それらのコドンに対応するイソロイシン(Ile)，アスパラギン(Asn)，リジン(Lys)，フェニルアラニン(Phe)などの出現頻度も高くなる．実際に，図 2 の解析に使用したデータセットについて，2 種の生物における各アミノ酸の頻度に順位をつけ，2 種間におけるスピアマンの順位相関係数 ρ を計算すると，アピコンプレクサ類に属する 4 種の生物と *Euglena longa* および *Helicosporidium* sp. のアミノ酸組成には，それぞれ他の生物よりも強い相関関係が認められた(ρ =0.77-0.89)．この結果は，これら 6 種の生物において Ile や Asn，Lys，Phe などの出現頻度が他の生物よりも高いことに起因する．従って図 2 の系統解析では，進化的距離の大きく離れたアピコンプレクサ類とユーグレナ藻類および *Helicosporidium* sp. が近縁関係にあるとするアーティファクトが，アミノ酸組成の類似性というバイアスによって強力に誘導されていると判断することができる．

上述のように塩基・アミノ酸組成の不均一性(あるいは特定の配列間における類似性)は一般的な分子系統解析の結果に大きく影響する問題である．これらのバイアスが均一な置換モデルを使用した分子系統解析の精度にどの程度悪影響を及ぼすかを定量的に評価することは重要な課題であるが，そのためには実配列データに基づく解析の他にシミュレーション実験による検証が不可欠である．何故なら，実配列に基づく系統解析では我々には「生物の真の系統関係を示す系統樹」を知る術がなく，系統解析の精度(= 真の系統樹の選択率)を正確に評価することが困難だからである．

塩基組成，特に AT 含量の不均一性が，塩基組成の均一性を仮定した置換モデルを使用する分子系統解析に与える影響については，既にいくつかのシミュレーション実験の結果が報告されている．これらはいずれも，タンパク質をコードしない塩基配列データの解析を想定したものである．Jermiin et al. (2004) や Ho and Jermiin (2004) では，図 1 のような 4 配列からなるモデル系統樹を用意し，モデル系統樹の樹形およびそれぞれの枝長に従って末端の仮想的塩基配列を生成し実験を行っている．これらの先行研究では，配列の生成に既存の塩基置換モデルを使用し，モデル系統樹の各枝に頻度パラメータ π を含む固有のパラメータを割り当て，パラメータの不均一な塩基置換モデルを適用することで，末端配列において実現される AT 含量に意図的に不均一性が生じるようにしている．このように不均一な置換プロセスに基づく配列進化を想定したシミュレーションによって生成された配列を，均一な塩基置換モデルを使用する一般的な分子系統解析に供することで，モデル系統樹（＝真の系統樹）の再現率を定量的に評価することが可能となったのである．

一方，タンパク質コーディング遺伝子配列（コドン配列）において塩基・アミノ酸組成の不均一性が分子系統解析の精度にどの程度影響するかという点も，同様に検証されるべき問題である．この検証を行うためには，まずコドン配列の進化を想定したコドン置換モデルによって配列生成を行うことが必要であり，その上で「配列間の各コドン座位における塩基組成の不均一性（コドン使用頻度の不均一性）」という点を想定した不均一モデルでのシミュレーションを行うことが求められる．しかしながら，それを実現する配列生成ソフトウェアは近年まで存在しなかったため，この問題の検証はこれまで詳細に行われていなかった．そこで今回筆者らは，近年になり開発されたソフトウェアである BppML および BppSeqGen (Dutheil and Boussau, 2008) を用いてこの問題を検証した．まず，先に述べた真核生物色素体およびシアノバクテリア由来の遺伝子配列データをサンプルとし，アピコプラストが紅藻起源色素体と近縁であるというモデル系統樹をもとに，不均一なコドン置換モデルを用いてタンパク質コーディング遺伝子の仮想的配列を生成した．こうして得られた配列では，各コドン座位における AT 含量や，塩基配列翻訳後のアミノ酸組成の不均一性を実現することができた．さらに，このデータを均一な置換モデルに基づく分子系統解析に供すことにより，その精度を定量的に評価し，均一なモデルが正しい系統樹を復元できないことをシミュレーション研究により初めて明確に示した．本稿ではこれらのデータ解析の成果を研究速報として報告する．

2. データ解析の方法

2.1 シミュレーションのパラメータ推定のためのサンプルデータとモデル系統樹の準備

仮想的配列をシミュレーションにより生成するためには，均一モデル・不均一モデルにかかわらず，配列進化を示すモデル系統樹を用意し，モデル系統樹の各枝長および配列生成に使用する置換モデルのパラメータを指定する必要がある．これらの値には，実配列データを使った系統解析から得られた推定値を使用するのが妥当である．本研究ではパラメータ推定のためのサンプルとして，図 2 の系統解析に用いた真核生物の色素体およびシアノバクテリアの 7 遺伝子（*rpl14*, *rpl16*, *rps3*, *rps11*, *rps12*, *rpoB*, *tufA*）の配列データを使用した．配列データの作成方法について以下に記述する．

まず，シアノバクテリア 2 種，紅藻類 3 種，珪藻類 2 種，緑藻類 3 種，ユーグレナ藻類 2 種，*Helicosporidium* sp. アピコンプレクサ類 4 種の計 17 種について，7 遺伝子それぞれの cDNA 配列およびアミノ酸配列を NCBI データベースより取得した．次に各遺伝子について，アミノ酸配列に基づくアライメントを行った．アライメントには MAFFT (Katoh et al., 2005) を用い，自動アライメントの後手動アライメントによる精査を行った．整理されたアライメントデータよりアライメントに曖昧さを伴わない座位を選択し，1,279 アミノ酸座位の系統解析用配列デー

タを作成した．さらに，このアミノ酸配列のアライメントに基づいて cDNA 配列のアライメントおよび座位選択を行い，最終的に 3,837 塩基の塩基配列(1,279 コドンのコドン配列)データを作成した．アミノ酸配列のアライメントに基づく cDNA 配列のアライメントには，PAL2NAL (Suyama et al., 2006) を使用した．

この手順で得られた塩基配列およびアミノ酸配列を用いた分子系統解析では，アピコンプレクサ類がユーグレナ藻類および *Helicosporidium* sp. と近縁であるというアーティファクトが誘導される(図 2)．そこで，本研究ではまず Janouškovec et al. (2010) などの過去研究より得られた知見に従い，「アピコンプレクサ類は珪藻類と最も近縁である」という制約条件の下で最尤系統樹を探索した．次にこれをモデル系統樹とし，cDNA 配列(コドン配列)を用いて，次節で述べるコドン置換モデル(YN98 モデル) (Yang and Nielsen, 1998)によりシミュレーションのための各種パラメータを推定した．

2.2 均一/不均一なコドン置換モデル

本節では今回の解析で用いた，コドン間の置換に関する確率モデルである YN98 モデルについて，均一な場合と不均一な場合に分けて簡潔に説明する．一つのコドンは第 1，2，3 座位の DNA 塩基の組み合わせによって表現されるためその総数は $4^3 = 64$ 個であるが，標準遺伝暗号 (Standard Genetic Code)に従うタンパク質コーディング遺伝子では TAG，TAA，TGA の 3 種類のコドンは終始コドンとして扱われるため，YN98 モデルでは実際には 61 個のコドン間での置換を考える．

まず，系統樹の各枝で配列の進化が均一な YN98 モデルに従う場合を考える．均一な YN98 モデルでは，全ての枝でコドン i から j への瞬間置換速度は次のように表現される．

$$q_{ij} = \begin{cases} 0 & \text{塩基置換が 2 回以上起こる} \\ u\pi_j & \text{同義置換かつトランスバージョン} \\ u\kappa\pi_j & \text{同義置換かつトランジション} \\ u\omega\pi_j & \text{非同義置換かつトランスバージョン} \\ u\omega\kappa\pi_j & \text{非同義置換かつトランジション} \end{cases} \tag{2.1}$$

ここで，π_j はコドン j の頻度を示す．なお，本研究で扱う YN98 モデルでは各コドン座位における DNA 塩基の頻度 $\pi_A, \pi_T, \pi_G, \pi_C$ をそれぞれ配列データより推定し，これらの値を用いて各コドン頻度 π_j を近似的に計算した．κ はトランジション(purine (A, G) あるいは pyrimidine (C, T) 同士の置換)のトランスバージョン(purine，pyrimidine 間の置換)に対する相対比を，ω は非同義置換と同義置換の速度比を表す．また，u は 1 単位時間に起こるコドン置換が 1 になるよう標準化するパラメータであり，π_j, κ, ω が決まると自動的に決定される．なお，コドン i とコドン j が等しい場合には $q_{ij} = -\sum_{j:j\neq i} q_{ij}$ とし，61×61 の置換モデル Q を定義する．すると，系統樹上の各枝 t_n に対応する遷移確率は次のようにして計算される．

$$P(t_n) = e^{t_n Q} \tag{2.2}$$

こうして得られる枝ごとの遷移確率より，系統樹の尤度を計算することが可能である．再度，図 1 に示したような 4 本の配列 A，B，C，D からなる系統樹の尤度を計算する場合を考える．配列 A，B，C，D の第 h 番目の座位においてコドン p，q，r，s が観測されたとし，配列 A，B および配列 C，D の祖先配列におけるコドンをそれぞれ i，j と仮定すると，この座位の尤度は

$$L(\theta|X_h) = f(X_h|\theta) = \sum_{i=1}^{61} \pi_i p_{ip}(t_1) p_{iq}(t_2) \sum_{j=1}^{61} p_{ij}(t_5 + t_6) p_{jr}(t_3) p_{js}(t_4) \tag{2.3}$$

によって表現される．コドン配列の全座位の尤度は式 1.2 より算出される．

次に，系統樹の各枝での配列進化が不均一な YN98 モデルに従う場合を考える．この場合，系統樹の m 番目の枝 t_m において，コドン i からコドン j への瞬間置換速度は次のように表現される．

$$q_{ij}^{(m)} = \begin{cases} 0 & \text{塩基置換が 2 回以上起こる} \\ u^{(m)}\pi_j^{(m)} & \text{同義置換かつトランスバージョン} \\ u^{(m)}\kappa^{(m)}\pi_j^{(m)} & \text{同義置換かつトランジション} \\ u^{(m)}\omega^{(m)}\pi_j^{(m)} & \text{非同義置換かつトランスバージョン} \\ u^{(m)}\omega^{(m)}\kappa^{(m)}\pi_j^{(m)} & \text{非同義置換かつトランジション} \end{cases} \tag{2.4}$$

均一な YN98 モデルの場合と同様に，$\pi_j^{(m)}$,$\kappa^{(m)}$,$\omega^{(m)}$ はそれぞれモデルごとのコドン出現頻度，トランジション/トランスバージョン比，非同義/同義置換の速度比を示す．系統樹の各枝 t_m に対応する遷移確率は転移行列 $Q^{(m)}$ を用いて以下のように計算される．

$$P(t_m) = e^{t_m Q^{(m)}} \tag{2.5}$$

枝ごとの遷移確率を求めることで系統樹の尤度を計算することが可能であるが，その際注意すべきことは，均一モデルとは異なりコドン置換が不可逆的なモデルによって生じるということである（Boussau and Gouy, 2006）．すなわち，ある枝 t においてコドン x,y 間の置換を考えるとき，$\pi_x p_{xy}(t) \neq \pi_y p_{yx}(t)$ である．従って，不均一モデルを使って尤度を計算する際には，系統樹の根（root）を指定する必要がある．再度簡単のため，図 1 の系統樹を考える．系統樹の各枝 $t_1 \sim t_6$ には，それぞれ固有の YN98 モデル $Q^{(1)} \sim Q^{(6)}$ が割り当てられるものとする．また，根（R）における祖先配列を X とし，配列 A，B，C，D の第 h 番目の座位においてコドン p，q，r，s が観測されたとすると，この座位の尤度は

$$\begin{aligned} L(\theta|X_h) = f(X_h|\theta) = & \sum_{x=1}^{61} \pi_x \sum_{i=1}^{61} p_{xi}^{(5)}(t_5) p_{ip}^{(1)}(t_1) p_{iq}^{(2)}(t_2) \\ & \times \sum_{j=1}^{61} p_{xj}^{(6)}(t_6) p_{jr}^{(3)}(t_3) p_{js}^{(4)}(t_4) \end{aligned} \tag{2.6}$$

によって表現される（計算式は Boussau and Gouy, 2006 を参照）．なお，全座位の情報より計算される系統樹の尤度は均一モデルの場合と同様に算出される（式（1.2））．

本研究では，真核生物色素体およびシアノバクテリア由来の 7 遺伝子配列データおよび図 3（A），（B）に示したモデル系統樹に基づき，均一/不均一な YN98 モデルのもとでの系統樹の枝長および各種パラメータを最尤推定した．ただし，不均一な YN98 モデルを用いたパラメータ推定では，系統樹のひとつひとつの枝に異なる置換モデルを当てはめるのではなく，系統樹上で 7 つのモデル変化点（$M_1 \sim M_7$）を指定し，変化点以降の枝に異なる置換モデルを当てはめることで 7 つの不均一な置換モデルによって配列の進化を表現するようにした（図 3（B））．図 3（B）では，例えば $t_1 \sim t_3$ では共通の置換モデル $Q^{(1)}$ が当てはめられるが，$t_4 \sim t_8$ ではそれとは異なる置換モデル $Q^{(2)}$ が当てはめられる．なお，図 3 では緑藻は単系統とならないため，変化点 M_3 を 2 か所に設定し，それぞれに同一の置換モデルを当てはめることとした．また，座位間の進化速度の不均一性は確率の等しい 4 個の速度クラスに分割される離散 Γ 分布（Yang, 1994）によって近似されるものとした（YN98 + Γ モデル）．なお，Γ 分布の形状母数 α は系統樹の枝間で不変であるものとした．以上の均一/不均一な YN98 モデルによるモデル系統樹の枝長および各種パラメータの推定には BppML（Dutheil and Boussau, 2008）を使用した．

2.3 配列生成およびシミュレーション配列の系統解析

図 3（A），（B）のモデル系統樹について均一/不均一な YN98 + Γ モデルによって推定された

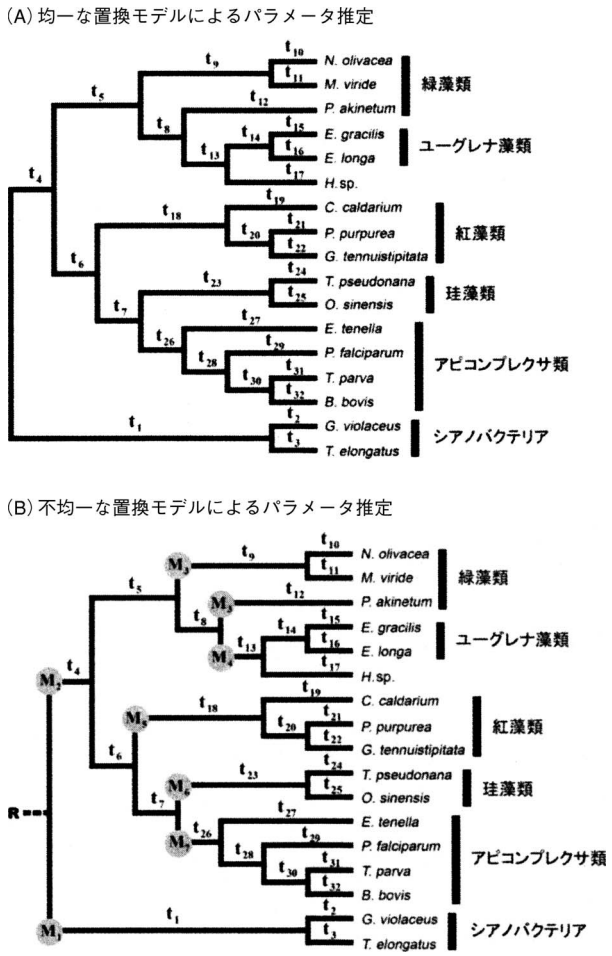

図 3. シミュレーションのためのパラメータ推定．(A) 均一な置換モデル，(B) 不均一な置換モデルそれぞれを使用した模式図を示す．t は系統樹の各枝，M，R，はそれぞれ不均一なモデルを適用する際必要となるモデルの変化点及び根 (root) の位置を示す．

枝長および各種パラメータ (π_j, κ, ω, α) を用いて，モンテカルロシミュレーション法による仮想的配列の生成を行った．配列生成には BppSeqGen (Dutheil and Boussau, 2008) を使用し，実データと同じ 1,279 コドン (3,837 塩基) の配列を生成した．不均一な置換モデルを用いた配列生成では，モデルの変化点 ($M_1 \sim M_7$) および根 (R) はパラメータ推定の際と同じ位置を指定した (図 3 (B))．なお，均一モデルによる配列生成では根の位置は指定せず，祖先的配列はランダムな位置で生成されるものとした．配列生成は 100 反復行い，100 個のデータセットを作成した．

次に，均一/不均一モデルによって生成された配列データを均一な置換モデルを用いた分子系統解析に供した．この分子系統解析では①コドンの全座位を用いた 3,837 塩基の配列データと塩基置換モデルに基づく解析，②コドン配列を標準遺伝暗号に従って翻訳したアミノ酸配列とアミノ置換モデルに基づく解析の 2 つを行った．①②のデータセットの解析には GTR + Γ

(Tavaré, 1986)，LG + Γ (Le and Gascuel, 2008)を用い，RAxML Ver. 7.2.8 (Stamatakis et al., 2008)を使用し，1 本の初期系統樹から発見的探索により最尤系統樹を推測した．最終的に塩基・アミノ酸配列それぞれ 100 個のデータセットより得られた最尤系統樹より合意系統樹を作成し，均一/不均一なモデルによる解析におけるモデル系統樹の各ノードの復元率を計算した．

3. 結果

3.1 シミュレーション配列における各コドン座位の AT 含量の不均一性

均一な YN98 モデルに基づく配列生成では，系統樹上の全ての配列進化は同一のパラメータ π_j, κ, ω に従う．そのため，モデル系統樹の末端配列では，全ての配列間で各コドン座位における AT 含量はほぼ一定になると予想される．実際に，均一な YN98 モデルを使用したシミュレーションにより生成された配列 100 個について各コドン座位の AT 含量を計算すると，アピコンプレクサ類を示す仮想的配列(グループ①)，ユーグレナ藻類・*Helicosporidium* sp. を示す仮想的配列(グループ②)およびそれ以外の生物を示す仮想的配列(グループ③)では，グループ内での平均的な AT 含量には殆ど差はみられなかった(表 1)．

一方，不均一な YN98 モデルによるシミュレーションでは系統樹上で置換モデルの各パラメータが変化するため，末端配列では各コドン座位の AT 含量に不均一性が生じることが予想される．実際に，不均一な YN98 モデルにより生成された配列 100 個について均一モデルによるシミュレーションと同様に各コドン座位の AT 含量を計算すると，グループ①および②の配列はグループ③の配列よりも各コドン座位で AT 含量が大きくなるという結果となった(表 1)．また，グループ①の配列はグループ②の配列よりも全ての座位で平均 AT 含量が大きくなった．これらの結果は，実配列におけるグループ①②③間での各コドン座位の AT 含量差とも矛盾しないものである(図 2)．

さらに，グループ①と②，①と③，②と③それぞれの間の AT 含量差の平均が，均一なモデルによるシミュレーションの場合と不均一なモデルによるシミュレーションの場合で有意に異なるかどうかを検定するために，それぞれのシミュレーション 100 反復の標本を用いて，平均の差の検定(t 検定)を行った．その結果，全てのコドン座位における 2 グループ間の組み合わせすべてについて，不均一モデルによるシミュレーションでの AT 含量差の平均は均一なモデルによるシミュレーションでの AT 含量差の平均とは有意に異なることが確認された(表 1)．

3.2 モデル系統樹の再現率

均一モデルにより生成された配列では全ての配列が同一の置換モデルに従って進化するため，均一な置換モデルを使った分子系統解析でも極端なモデル不整合は生じない．本研究の解析でも，均一モデルによるシミュレーションでは，塩基，アミノ酸どちらの配列に基づく解析であってもモデル系統樹(＝真の系統樹)と同じ樹形が 100 反復中ほぼ 100%最尤系統樹として選択された(図 4 (A))．また，実配列を使用した解析において誘導された「アピコンプレクサ類が緑藻および緑藻起源二次葉緑体をもつ生物と近縁関係にある」というアーティファクトの樹形は，均一モデルによるシミュレーションでは全く選択されなかった．

その一方で不均一モデルによるシミュレーションでは，各コドン座位における AT 含量に不均一性が生じるため，シミュレーション配列の塩基・アミノ酸組成にも不均一性が生じることになる．そのため，均一な置換モデルを用いた系統解析ではモデルの不整合が生じ，真の系統樹とは異なるアーティファクトが誘導されることが予想される．事実，不均一モデルによるシミュレーションでは，「アピコンプレクサ類が珪藻類および紅藻類と近縁である」ことを示す真の系統樹のノードの復元率が塩基配列に基づく解析では 3%，アミノ酸配列に基づく解析では 8%と極端に低くなった(図 4 (B))．他方，「アピコンプレクサ類がユーグレナ藻類・*Helicosporidium*

表 1. シミュレーション配列における各コドン座位での AT 含量差.

コドン第1座位

サンプル配列群(1, 2)	$ave_{100reps}\{\Delta[ave\ AT^{(1)}, ave\ AT^{(2)}]\}^{Homo}$ (%)	$ave_{100reps}\{\Delta[ave\ AT^{(1)}, ave\ AT^{(2)}]\}^{NH}$ (%)	t検定(p 値)
(グループ①, グループ③)	0.022	1.578	<< 0.01
(グループ②, グループ③)	-0.002	0.857	<< 0.01
(グループ①, グループ②)	0.019	0.707	<< 0.01

コドン第2座位

サンプル配列群(1, 2)	$ave_{100reps}\{\Delta[ave\ AT^{(1)}, ave\ AT^{(2)}]\}^{Homo}$ (%)	$ave_{100reps}\{\Delta[ave\ AT^{(1)}, ave\ AT^{(2)}]\}^{NH}$ (%)	t検定(p 値)
(グループ①, グループ③)	-0.004	0.572	<< 0.01
(グループ②, グループ③)	0.001	0.194	<< 0.01
(グループ①, グループ②)	-0.008	0.374	<< 0.01

コドン第3座位

サンプル配列群(1, 2)	$ave_{100reps}\{\Delta[ave\ AT^{(1)}, ave\ AT^{(2)}]\}^{Homo}$ (%)	$ave_{100reps}\{\Delta[ave\ AT^{(1)}, ave\ AT^{(2)}]\}^{NH}$ (%)	t検定(p 値)
(グループ①, グループ③)	-0.064	5.262	<< 0.01
(グループ②, グループ③)	-0.014	4.588	<< 0.01
(グループ①, グループ②)	-0.050	0.687	<< 0.01

グループ①：(*E. tenella, P.falciparum, T.parva, B.bovis*), グループ②：(*E.gracilis, E.longa, H.* sp), グループ③：グループ①②以外の生物種

$\Delta[ave\ AT^{(1)}, ave\ AT^{(2)}]$はサンプル群1, 2それぞれにおけるAT含量の平均値の差を示す

$ave_{100reps}$ はシミュレーション100反復から計算された平均値を示す

Homo：均一(Homogeneous)モデルによるシミュレーション, NH：不均一(non-Homogeneous)モデルによるシミュレーション

第1,2座位でのt検定はWelchのt検定, 第3座位でのt検定はStudentのt検定によるp 値を示す

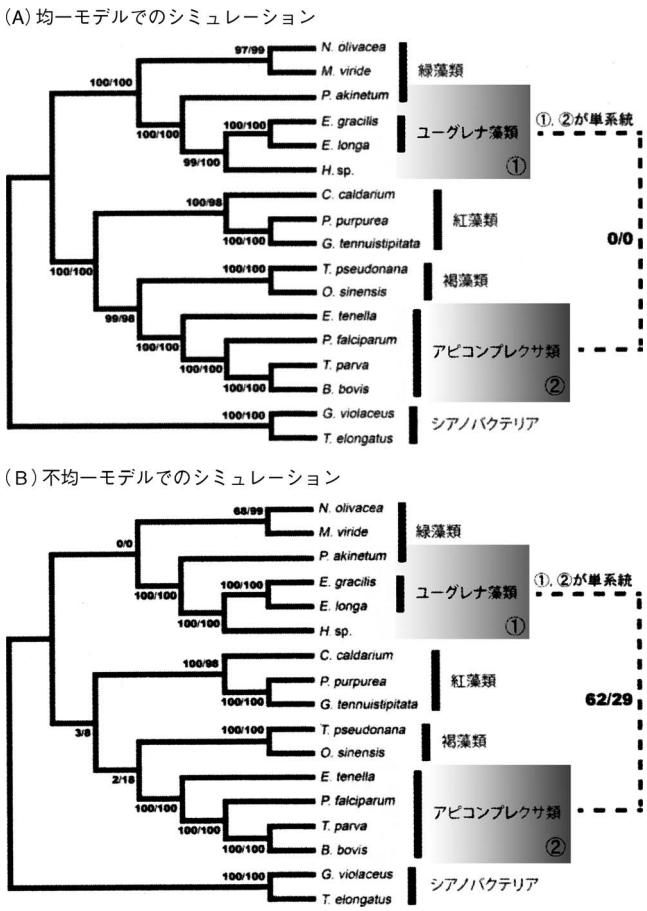

図 4. シミュレーション実験におけるモデル系統樹及びアーティファクトの選択率. (A), (B) の樹形はモデル系統樹 (図 3) に一致する. 各ノードの値は左から①コドンの全座位を使用した塩基配列に基づく解析, ②コドン配列を翻訳したアミノ酸配列に基づく解析それぞれについて, 100 反復の実験における復元率を示す.

sp. および緑藻類 *Pseudenoclonium akinetum* と近縁である」ことを示すノードの選択率は, 塩基配列による解析では 62%, アミノ酸配列による解析では 29% と均一モデルによるシミュレーションの場合に比べ大きく上昇した.

4. 考察

タンパク質コーディング遺伝子において, 塩基・アミノ酸組成の不均一性は実配列を使った系統解析では頻繁にみられる問題であり, そのような配列データに対し進化プロセスの均一性を仮定する置換モデルを適用した分子系統解析では, モデルの不整合により真の系統樹とは異なるアーティファクトが誘導される危険性がある. 本研究は, 各コドン座位における AT 含量やアミノ酸組成に極端な不均一性がみられる実配列データをサンプルとし, 不均一なコドン置換モデルを利用したシミュレーションを行うことで, タンパク質コーディング遺伝子における配列組成の不均一性が一般的な分子系統解析の精度に与える影響を定量的に評価することに初

めて挑戦した研究である．

本研究で行った不均一モデルによるシミュレーションでは，モデル系統樹上でアピコンプレクサ類あるいはユーグレナ藻類・*Helicosporidium* sp. が分岐する直前の枝で置換モデルが変化する(図 3 (B)の M_4，M_7)．変化後の置換モデルには，i 番目のコドン座位における塩基組成 π_A^i, π_T^i, π_G^i, π_C^i が新たに指定されるが，ここでの π_A^i, π_T^i は実配列より推定された値をもとに他の枝に指定される値よりも大きなものが指定される．そのため，アピコンプレクサ類あるいはユーグレナ藻類・*Helicosporidium* sp. のシミュレーション配列では，各コドン座位の AT 含量が特異的に上昇することになる．必然的に，これらの配列全体の AT 含量も上昇することになるため，塩基配列に基づく均一な置換モデルによる分子系統解析ではモデルの不整合により，高い AT 含量をもつアピコンプレクサ類とユーグレナ藻類・*Helicosporidium* sp. が近縁関係にあるとみなすアーティファクトが強力に誘導されることとなった(図 4 (B))．

また，各コドン座位の AT 含量に偏りがある場合にはコドン使用頻度にも偏りが生じ，各座位に A あるいは T を多用するコドン(AAA，TTT など)が頻繁に使用されることになる．その結果，塩基配列翻訳後のアミノ酸配列においても，それらのコドンに対応する Ile，Asn，Lys，Phe などの出現頻度が上昇することになる．不均一モデルによるシミュレーションでは，アピコンプレクサ類あるいはユーグレナ藻類・*Helicosporidium* sp. ではこれらのアミノ酸の出現頻度が他の生物よりも大きくなることで，均一なアミノ酸置換モデルを用いた分子系統解析でもモデルの不整合が起こり，塩基配列に基づく解析と同様のアーティファクトが誘導された(図 4 (B))．なお，タンパク質コーディング遺伝子ではコドンの第 3 座位の進化速度が最も大きくなるため，AT 含量に偏りがある場合，その偏りは第 3 座位において最も極端になることが分かっている(Carlini et al., 2001)．本研究で行った不均一モデルによるシミュレーションでも，アピコンプレクサ類，ユーグレナ藻類・*Helicosporidium* sp. における AT 含量の偏りは第 3 座位において最も大きくなった(表 1)．ここで，コドンの第 3 座位における塩基置換は多くの場合アミノ酸の変化を伴わない同義置換であるため，塩基配列をアミノ酸に翻訳して解析に用いることで，コドン第 3 座位における塩基組成の不均一性を無視することが可能である(Hashimoto et al., 1994)．しかしながら，コドンの第 1，第 2 座位においても AT 含量に極端な偏りが生じる場合には，翻訳後のアミノ酸組成にも不均一性が生じるため，アミノ酸配列を用いた系統解析であってもモデル不整合は免れないことが予想される．本研究の不均一モデルによるシミュレーションでは，アピコンプレクサ類あるいはユーグレナ藻類・*Helicosporidium* sp. ではコドンの第 1，第 2 座位においても AT 含量に極端な偏りが生じている(表 1)．そのため，アミノ酸配列を用いた分子系統解析であっても依然としてアーティファクトそのものは誘導されており，モデル系統樹全体の再現率にも効果的な改善がみられるわけではない(図 4 (B))．

なお，本研究のシミュレーションでは，各コドン座位の AT 含量の偏りは実配列データから観測される値よりも小さい範囲に収まっている．表 2 では，アピコンプレクサ類(グループ①)あるいはユーグレナ藻類・*Helicosporidium* sp. (グループ②)と，それ以外の生物(グループ③)について，各コドン座位における AT 含量のグループ間での平均値の差を計算し，実配列からの観測値とシミュレーションにおける実現値との比較を行ったが，実配列からの観測値はシミュレーションにおける実現値に比べ，より極端な値を示した．特に，グループ①②ではコドン第 1，第 2 座位における AT 含量もグループ③に比べ大きく上昇しており，翻訳後のアミノ酸配列の組成にも変化を伴うレベルで AT 含量の上昇が生じていることを示している．このことから，実配列データでは塩基・アミノ酸組成の不均一性がより強力なバイアスとして存在し，分子系統解析の精度により重大な影響を及ぼしていることが示唆される．実際に，真核生物色素体およびシアノバクテリア由来の 7 遺伝配列を用いた分子系統解析では，塩基配列・アミノ酸配列いずれに基づく解析にかかわらず，アーティファクトの選択率，すなわちアピコンプレク

表 2. 実配列およびシミュレーション配列における各コドン座位での AT 含量差.

コドン第1座位

サンプル配列群(1, 2)	Δ [ave $AT^{(1)}$, ave $AT^{(2)}$]^{Original} (%)	$ave_{100reps}$ { Δ[ave $AT^{(1)}$, ave $AT^{(2)}$] }^{NH} ± 2*SD (%)
(グループ①, グループ③)	3.978	1.578 ± 0.5
(グループ②, グループ③)	8.107	0.857 ± 0.4
(グループ①, グループ②)	4.129	0.707 ± 0.5

コドン第2座位

サンプル配列群(1, 2)	Δ [ave $AT^{(1)}$, ave $AT^{(2)}$]^{Original} (%)	$ave_{100reps}$ { Δ[ave $AT^{(1)}$, ave $AT^{(2)}$] }^{NH} ± 2*SD (%)
(グループ①, グループ③)	1.492	0.572 ± 0.3
(グループ②, グループ③)	3.913	0.194 ± 0.3
(グループ①, グループ②)	2.422	0.374 ± 0.4

コドン第3座位

サンプル配列群(1, 2)	Δ [ave $AT^{(1)}$, ave $AT^{(2)}$]^{Original} (%)	$ave_{100reps}$ { Δ[ave $AT^{(1)}$, ave $AT^{(2)}$] }^{NH} ± 2*SD (%)
(グループ①, グループ③)	5.353	5.262 ± 0.5
(グループ②, グループ③)	6.813	4.588 ± 0.6
(グループ①, グループ②)	1.459	0.687 ± 0.7

グループ①：(*E. tenella, P.falciparum, T.parva, B.bovis*), グループ②：(*E.gracilis, E.longa, H.* sp), グループ③：グループ①②以外の生物
Δ[ave $AT^{(1)}$, ave $AT^{(2)}$]はサンプル群1, 2それぞれにおけるAT含量の平均値の差を示す
$ave_{100reps}$ はシミュレーション100反復から計算された平均値を示す
Original：色素体由来7遺伝子(*rpl14, rpl16, rpoB, rps3, rps12, tufA*)からの観測値, NH：不均一モデルによるシミュレーション

サ類とユーグレナ藻類・*Helicosporidium* sp. を結びつける内部枝のブートストラップ値は，シミュレーション実験における選択率よりも高い値を示している(図 2).

5. おわりに

本研究のシミュレーション実験により，タンパク質コーディング遺伝子における組成の不均一性が均一性を仮定する置換モデルを使った分子系統解析に与える影響を定量的に評価できた．本研究で使用した真核生物色素体およびシアノバクテリア由来の遺伝子配列データは一つのサンプルに過ぎないが，他の実データ解析においても，本研究で想定された程度の塩基・アミノ酸組成の不均一性が確認された場合には，分子系統解析には少なからずバイアスが働いているものとして解析結果を慎重に精査するべきであろう．なお近年になって，本研究で使用した BppML や，nhPhyloBayes (Blanquart and Lartillot, 2008) など，系統間でパラメータの不均一な置換モデルを使用して系統樹の推測を行うことができるソフトウェアが開発されている．このようなソフトウェアを使用すれば，配列組成が不均一なデータに基づく分子系統解析の精度は飛躍的に向上することが期待されるが，その有効性をさらに詳細に検証するためには，本研究で用いたような手法に基づくシミュレーション実験が必要であろう．

参 考 文 献

Blanquart, S. and Lartillot, N. (2008). A site- and time-heterogeneous model of amino-acid replacement, *Molecular Biology and Evolution*, **25**, 842–858.

Boussau, B. and Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution, *Systematic Biology*, **55**, 756–768.

Carlini, D. B., Chen, Y. and Stephan, W. (2001). The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the Drosophilid

alcohol dehydrogenase genes, *Adh* and *Adhr*, *Genetics*, **159**, 623–633.

Chang, B. S. W. and Campbell, D. L. (2000). Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences, *Molecular Biology and Evolution*, **17**, 1220–1231.

Dutheil, J. and Boussau, B. (2008). Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs, *BMC Evolutionary Biology*, 8: 255. doi: 10.1186/1471-2148-8-255.

Hashimoto, T., Nakamura, Y., Nakamura, F., Shirakura, T., Adachi, J., Goto, N., Okamoto, K. and Hasegawa, M. (1994). Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia, Molecular Biology and Evolution*, **12**, 782–793.

Ho, S. Y. W. and Jermiin, L. S. (2004). Tracing the decay of the historical signal in biological sequence data, *Systematic Biology*, **53**, 623–637.

Janouškovec, J., Horak, A., Obornik, M., Lukes, J. and Keeling, P. J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids, *Proceedings of the National Academy of Science of the United States of America*, **107**, 10949–10954.

Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J. and Larkum, A. W. D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated, *Systematic Biology*, **53**, 637–643.

Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005). MAFFT version 5 : Improvement in accuracy of multiple sequence alignment, *Nucleic Acids Research*, **33**, 511–518.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix, *Molecular Biology and Evolution*, **25**, 1307–1320.

Lockhart, P. J., Steel, M. A, Hendy, M. D. and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution, *Molecular Biology and Evolution*, **11**, 605–612.

Mooers, A. and Holmes, E. C. (2000). The evolution of base composition and phylogenetic inference, *Trends in Ecology and Evolution*, **15**, 365–369.

Moore, R. B., Obornik, M., Janouškovec, J., Chrudimsky, T., Vancova, M., Green, D. H., Wright, S. W., Davies, N. W., Bolch, C. J., Heimann, K., Slapeta, J., Hoegh-Guldberg, O., Logsdon, J. M. and Carter, D. A. (2008). A photosynthetic alveolate closely related to apicomplexan parasites, *Nature*, **451**, 959–963.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. and Reyes, A. (1999). Evolutionary genomics in Metazoa: The mitochondrial DNA as a model system, *Gene*, **238**, 195–209.

Stamatakis, A., Hoover, P. and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers, *Systematic Biology*, **57**(5), 758–771.

Stenico, M., Lloyd, A. T. and Sharp, P. M. (1994). Codon usage in Caenorhabditis elegans: Delineation of translational selection and mutational biases, *Nucleic Acids Research*, **22**, 2437–2446.

Suyama, M., Torrents, D. and Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Research*, **34**, 609–612.

Tarrio, R., Rodríguez-Trelles, F. and Ayala, F. J. (2001). Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae, *Molecular Biology and Evolution*, **18**, 1464–1473.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences, *Lecture on Mathematics in the Life Science*, **17**, 57–86.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods, *Journal of Molecular Evolution*, **39**, 306–314.

Yang, Z. and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals, *Journal of Molecular Evolution*, **46**, 409–418.

Proceedings of the Institute of Statistical Mathematics Vol. 60, No. 2, 289–303 (2012) 303

Assessment of the Performance of Phylogenetic Inference Based on Simulated Protein-coding Sequences with Significant Compositional Heterogeneity

Sohta A. Ishikawa and Tetsuo Hashimoto

Graduate School of Life and Environmental Sciences, University of Tsukuba

Phylogenetic analyses of molecular sequence data with commonly-used 'homogeneous' substitution models assume the stationarity of nucleotide or amino-acid composition across tree, but real world data sometimes violate the assumption. This report assesses how significantly the violation of compositional stationarity affects the performance of homogeneous model-based phylogenetic inference by using simulated protein-coding sequences. In order to estimate parameters for sequence simulation, we prepared a real-world sequence data set of seven plastid genome-encoded protein genes with adenine plus thymine content (AT content) in all the 1st, 2nd, and 3rd codon positions extraordinarily biased between species, and subjected it to a maximum-likelihood analysis for a given model tree to estimate the parameters. The analysis was carried out assuming a 'non-homogeneous' codon substitution model that can accommodate the heterogeneity of nucleotide composition in three codon positions across the tree. Using the parameters estimated and the model tree, we simulated protein-coding sequence data with compositional heterogeneity between species by the Monte-Carlo method. Finally, we tested the performance of homogeneous model-based phylogenetic analyses both at nucleotide and amino acid sequence levels for recovering the model ('correct') tree. The results clearly demonstrated that both of the two analyses mostly failed to recover the correct tree but instead strongly favored artifactual trees attracted by the parallel compositional convergence between distantly-related species. This is de facto the first simulation study that assessed the appropriateness of applying homogeneous substitution models in phylogenetic analyses to protein-coding sequence data containing significant compositional heterogeneity.

Key words: Phylogenetic inference, maximum-likelihood analysis, compositional heterogeneity, codon substitution model, simulation, model misspecification.