統計数理(2008)
第 56 巻 第 1 号 19–35

特集「分子進化と統計科学」
[研究詳解]

ゲノム系統学的手法の応用と課題

―真獣類の起源に関する解析を例として―

西原 秀典[1] ・岡田 典弘[1] ・長谷川 政美[2]

(受付 2007 年 9 月 6 日；改訂 2007 年 12 月 21 日)

要　　旨

近年のゲノムプロジェクトの急速な進行とともに，様々な生物種に関して全ゲノム規模のデータを用いた系統樹推定(Phylogenomics)がおこなわれるようになってきた．配列データ量の増加が系統解析に有用であることは言うまでもないが，もし系統樹推定の際に仮定する進化モデルに偏りがあった場合，誤った結論を導いてしまうことがある．本稿では，大量の遺伝子配列を結合させたデータセットの解析(Concatenate model)ではおそらく誤りであろう系統仮説を強く支持したが，遺伝子ごとに異なる進化モデルを仮定した場合(Separate model)はその系統樹推定の偏りが激減するという極端な例を紹介する．本研究では 2,789 個の遺伝子配列(1 Mbp 以上)のデータセットを用い，真獣類の初期進化，すなわちアフリカ獣類，貧歯類，北方獣類の間の系統関係に関して最尤法を用いた解析をおこなった．その結果，従来の一般的な解析方法である塩基配列の Concatenate model ではアフリカ獣類と貧歯類の単系統性が 100%のブートストラップ(BP)値を伴って支持されたが，遺伝子間で異なる進化速度・進化パターンを仮定する Separate model ではその仮説がほとんど支持されなかった．この結果から，遺伝子配列データが膨大であっても全配列に対して同一の進化モデルを仮定してしまうと誤った結論を導くことになってしまうことがあり，それを避けるためには進化速度・進化パターンが遺伝子ごとに異なることを仮定する Separate model を適用すべきであることが示された．

キーワード： 真獣類の初期進化，分子系統樹の最尤推定，ゲノム系統学，Separate model.

1. はじめに

2001 年にヒトゲノムの概要配列が解読されたのをはじめとして，様々な生物種の全ゲノム DNA 解読プロジェクトが近年急速に進められるようになった．哺乳類では現在までに 7 種の全ゲノム配列が報告され，さらに 20 種以上のゲノムプロジェクトが進行中である．分子データを用いて生物種間の類縁関係を推定する分子系統学において，近年の全ゲノム配列データは解析手法に大きな影響を与えている．系統樹推定においては 1 つあるいは少数の遺伝子配列を解析するのが一般的な方法であるが，Taxa 数の少なさなどが原因で推定に誤りが生じる場合がある．しかし大量の遺伝子データはこうした標本誤差を小さくするため，ポストゲノム時代に入ると全ゲノム情報を利用した系統樹推定が徐々におこなわれるようになってきた．こうし

[1] 東京工業大学大学院 生命理工学研究科：〒 226–8501 横浜市緑区長津田町 4259–B–21
[2] 復旦大学 生命科学学院：220 Handan Road, Shanghai 200433, China

た全ゲノム規模の比較解析によって信頼性の高い系統樹が推定され，さらにそれを基盤としてゲノム構造や機能がどのような進化過程を経たのかを推定することが可能となる．こうしたアプローチはゲノム系統学(Phylogenomics)と呼ばれ，ポストゲノム時代の新しい分野として確立しつつある(Rokas et al., 2003; Soltis et al., 2004; Delsuc et al., 2006)．しかしながら，もし系統樹推定において仮定する進化モデルに偏りがあった場合，誤った仮説を強く支持してしまうことがあることが最近報告されてきている(Blair et al., 2002; Phillips et al., 2004; Delsuc et al., 2005; Dopazo and Dopazo, 2005; Philippe et al., 2005a; Jeffroy et al., 2006)．したがって，大量のゲノムデータから信頼性の高い系統樹を推定するためには，こうした偏りを小さくする方法を確立することが必要不可欠である．現在，哺乳類においては様々な種のゲノムプロジェクトが急速に進行しており，そのデータが容易に手に入る状況にある．したがってゲノム系統学における諸問題を解析・評価するためには，こうした哺乳類のゲノムデータを利用して系統樹推定をおこなうことが最適であると考えられる．本研究では，真獣類の中でどのグループが最初に分岐したのかを解明するため，ゲノム系統学的手法を用いた解析をおこなった．

2. 真獣類の初期進化

哺乳類は現在 4 千種以上が知られており，陸上のみならず海や空にも進出していることから地球上で最も繁栄している動物であると言われている．現生の哺乳類は 20 目に分類されているが，その中でも単孔目や有袋目は比較的初期に分岐したグループである．残る 18 目に属する哺乳類はすべて完全な胎盤を持ち，真獣類(または有胎盤類)と呼ばれている．哺乳類の系統進化に関しては以前から数多くの形態学的，古生物学的，分子系統学的研究がおこなわれてきたが，1990 年代前半までは分子系統解析に用いる情報量の少なさなどが原因で高次分類群間の関係はあまり解明されていなかった．しかし哺乳類の分子系統学はここ 10 年で急速に発展し，現在では目レベルの分類群間の系統関係はほとんど解明されたと言える(Madsen et al., 2001; Murphy et al., 2001a, 2001b; Kriegs et al., 2006; Nishihara et al., 2006)．結果的に，分子データから推定された系統樹はそれまで形態学的観点から提唱されてきた系統樹とは大きく異なり，従来の仮説を覆す部分が多かった(Cao et al., 2000)．分子系統学的結論では，現生の真獣類は大別してアフリカ獣類(Afrotheria)，貧歯類(Xenarthra)，北方獣類(Boreotheria)の 3 つの高次分類群に分けられる．アフリカ獣類は，ゾウ，海牛類(ジュゴン・マナティー)，ハイラックス，テンレック，キンモグラ，ツチブタ，ハネジネズミを含む 6 目から成る分類群である．アフリカ獣類は形態的に大きく多様化を遂げており形態学的にはその単系統性は認められていないが，分子系統学的観点からはその単系統性は間違いないとされている．またその多くの種がアフリカに生息していることからアフリカ獣類の起源はアフリカ大陸であり，かつてアフリカ大陸が他の大陸と分離して孤立していた時期にアフリカ獣類の多様化が進んだと考えられている．一方，貧歯類はアルマジロ，ナマケモノ，アリクイから成る単一の目であり，その起源は南米大陸であると考えられている．また北方獣類は 11 目から成り，ヒトを含む霊長目もこのグループに属する．北方獣類には海に進出したクジラや空に進出したコウモリなど多種多様な生物種が含まれ，非常に大きな分類群である．北方獣類はかつて北半球に存在したローラシア大陸(北米・ユーラシア大陸)が起源であると考えられている．

この 3 分類群間の系統関係，すなわち真獣類の中でどのグループが最初に分岐したのかという問題に関してはこれまでミトコンドリアや核遺伝子配列を用いて解析されてきたが，現在でも解明されないままである．さらに近年では複数の遺伝子配列を結合したデータセットが用いられているものの，この問題は解決されなかった(Waddell et al., 1999; Delsuc et al., 2002; Waddell and Shelley, 2003; Amrine-Madsen et al., 2003; Springer et al., 2004)．一方で，レトロポゾンの

挿入パターンに基づいた系統解析も最近おこなわれている．例えば Kriegs et al. (2006) は貧歯目が最初に分岐してアフリカ獣類と北方獣類が単系統であることを示すレトロポゾン挿入遺伝子座を 2 つ発見している．また Murphy et al. (2007) はアフリカ獣類と貧歯類の単系統性を示す遺伝子座を 2 つ報告しており，Kriegs et al. (2006) とは矛盾した結果となっている．しかしながら，これらはいずれも遺伝子座の数が少なく，レトロポゾン探索における ascertainment bias の可能性もあるため，この問題を解決する証拠としては弱い．

一方で，地質学的にはかつて南半球にはアフリカ大陸と南米大陸が繋がったゴンドワナ超大陸が存在し，それが約 1 億年前に分断されたとする仮説が提唱されている (Smith et al., 2004)．また真獣類の 3 グループが分岐したのもおよそ 1 億年前であると推定されている (Kumar and Hedges, 1998)．そのため仮にアフリカ獣類と貧歯類の単系統性が証明されれば，大陸の分断に伴って種分化が引き起こされた可能性も考えられる (Waddell et al., 1999)．したがって，この 3 グループ間の系統関係を明らかにすることは，真獣類の起源を解明するのみならず，大陸移動と哺乳類の分岐・移住の過程を解明するためにも必要なことである．本研究では，ゲノムデータベースから 2,789 遺伝子 (約 1 Mbp) の配列データセットを収集し，最尤法によってこの 3 グループ間の系統解析をおこなった．

3. 遺伝子配列データの取得

本研究では以下の 5 つの手順によって遺伝子配列データセットを収集した．(1) ヒトゲノムから 201 bp 以上のエキソン配列を収集，(2) 重複 (パラログ) 配列のデータを除去，(3) ヒトのエキソン配列のホモログをアフリカゾウおよびココノオビアルマジロのゲノムデータから検索，(4) 収集したエキソン配列のホモログを他の哺乳類ゲノムデータから取得，(5) 全配列のアライメントをおこない欠失サイトを除去．各手順の詳細を以下に記す．

本研究ではヒトゲノムをはじめ，ほとんどのデータを UCSC Genome Bioinformatics データベース (Hinrichs et al., 2006; http://genome.ucsc.edu/) から取得した．ヒトゲノムに関しては 2003 年に解読が完了し，全ゲノム配列データを染色体ごとに取得可能である．また UCSC データベースでは遺伝子の位置情報 (refFlat) も提供しており，それも取得した．なお，本研究ではヒトゲノムデータのバージョンとして hg17 を用いている．まずこの遺伝子の位置情報を参照し，ヒトゲノム配列データからタンパクをコードするエキソン配列をすべて抜き出した．この際，短いエキソン配列は後の BLAST 検索におけるホモログの単離が困難になるため長さが 201 bp 以上のエキソンに限定した．次に，得られたエキソン配列間で BLAST プログラムを用いた相同性検索をおこない，重複配列データを探索した．この際に自身の配列以外の配列データが 1×10^{-11} 以下の E-value を伴ってヒットした場合にパラログ配列であるとみなし，両配列をデータセットから除去した．この操作により 50,527 のエキソン配列が残り，これらはヒトゲノムにおいて単一コピーであると考えられる．

哺乳類の中でアフリカ獣類や貧歯類など比較的古くに分岐したグループのゲノムプロジェクトも進行中である．本研究では，アフリカ獣類の代表としてアフリカゾウ (*Loxodonta africana*)，貧歯類の代表としてココノオビアルマジロ (*Dasypus novemcinctus*) の 2x ショットガン配列データを DDBJ から取得し，解析に用いた．次に，BLAST プログラムを用いてヒトゲノムから収集したエキソン配列のホモログをこの 2 種のゲノムから探索した．その際に 1×10^{-11} 以下の E-value を伴ってヒットした配列のみを取得したが，この条件で 2 つの配列がヒットした場合はパラログの可能性があるとみなし，そのエキソン配列をデータセットから除去した．こうして 7,068 配列のホモログをヒト，ゾウ，アルマジロ間で収集した．

UCSC Genome Bioinformatics データベースでは，ヒトと各哺乳類との全ゲノムアライメン

トデータが取得できる．そこでこのアライメントデータを参照して 7 種の哺乳類ゲノムから各エキソン配列のオーソログを収集した．用いた種は，チンパンジー（*Pan troglodytes*，バージョンは panTro1），アカゲザル（*Macaca mulatta*，rheMac1），マウス（*Mus musculus*, mm7），ラット（*Rattus norvegicus*，rn3），イヌ（*Canis familiaris*，canFam2），ウシ（*Bos taurus*，bosTau1），オポッサム（*Monodelphis domestica*，monDom1）である．本研究では，オポッサムをアウトグループとして用いた．この過程で不明瞭な配列を除くために，1 種でも配列データが取得できない場合，またはエキソン配列内にストップコドンが検出された場合にはそのエキソンデータを除いた．これによって上記 10 種からそれぞれ 4,782 エキソン配列が収集され，アライメントをおこなった．全エキソン配列は一度 1 繋がりの配列として結合し，blastz (Schwartz et al., 2003)，および multiz (Blanchette et al., 2004) プログラムを用いてアライメントをおこなった．このプログラムでは系統情報を入力するため，北方獣類内の関係に関しては過去の文献（Murphy et al., 2001b; Nishihara et al., 2006）を参考にして既に解明されている系統関係を前提としたアライメントをおこなった．アフリカ獣類，貧歯類，北方獣類間の関係は多分岐とした．その後，アライメントされた配列データを各エキソンに分割し，さらに 1 つの遺伝子から複数のエキソンを用いている場合には遺伝子ごとにエキソンを統合した．さらに，もしある種においてコドン内で挿入・欠失が検出された場合，そのコドン配列を除いた．これにより得られた 3,148 遺伝子のデータセットは挿入・欠失の存在しないものとなったが，実在する遺伝子配列とは多少異なる場合がある．また，当初ヒトゲノムから取得したエキソン配列は 201 bp 以上であるが，BLAST 検索の際に非常に短い配列のみがヒットする場合がある．その結果得られた短い遺伝子配列を除去するため，120 bp 未満の遺伝子配列データをデータセットから除去した．

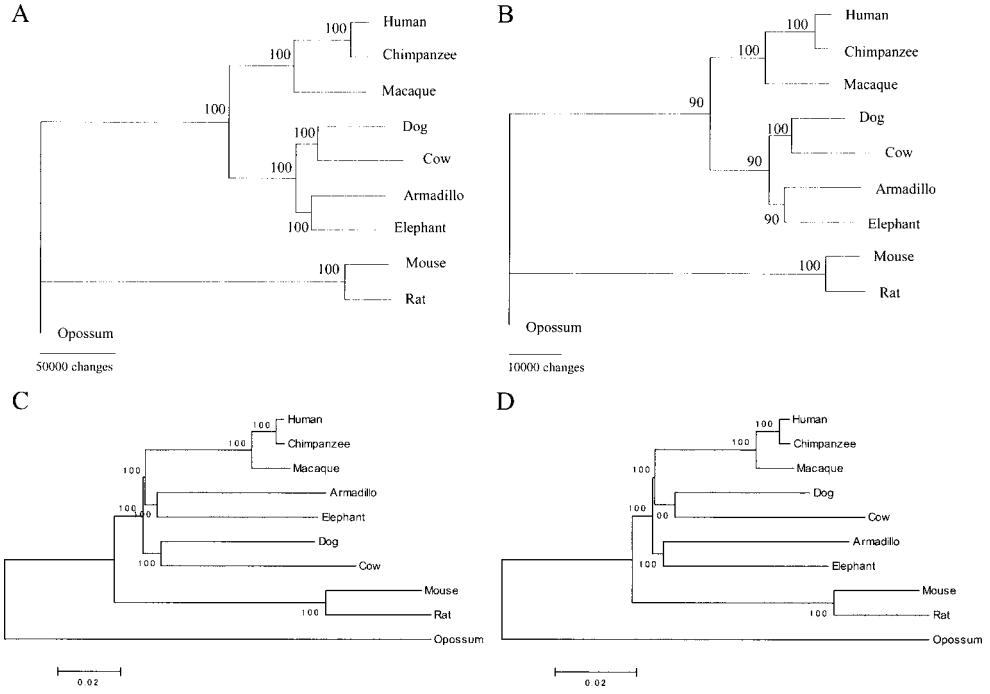

図 1. 全データセットを用いた MP および NJ 系統樹．(A) 塩基配列を用いた MP 系統樹，(B) アミノ酸配列を用いた MP 系統樹，(C) 塩基配列を用いた NJ 系統樹（Tamura-Nei model），(D) アミノ酸配列を用いた NJ 系統樹 (Poisson model)．

最終的に 2,789 遺伝子データセット(1,011,870 bp，337,290 コドン)を収集し以降の系統解析に用いた．これまでのゲノム系統学においては百数十 kbp のデータセットを用いた例が少数あるのみであったが，本研究で用いるデータセットは 1 Mbp 以上という量的に十分なものであり，また不明瞭なコドン配列やパラログ・偽遺伝子情報を可能な限り除いたという点において質的にも系統解析に適したものであると考えられる．

4. 近隣結合法および最節約法を用いた系統解析

収集したデータセットを用い，まず近隣結合(NJ)法および最節約(MP)法を用いた系統解析をおこなった．NJ 系統樹は MEGA3.0 を用いて Tamura-Nei モデルにより解析し，MP 系統樹は PAUP* 4.0 を用いて branch-and-bound によって最節約樹を探索した．その結果が図 1 である．いずれの系統樹においても真獣類の中でげっ歯類(マウス・ラット)が最初に分岐したという結果になった．しかしながらこれまでの研究結果からげっ歯類が北方獣類内部に含まれることは疑いようがないため，これは明らかにげっ歯類の進化速度が速いことが原因でおこった long branch attraction であると結論付けられる．一方，アミノ酸配列を用いた最尤系統樹においては北方獣類の単系統性は強く支持された(詳細は割愛する)．前述のようにこのデータセットはアライメントの際に北方獣類が単系統であるという系統情報を含めており，それにも関わらず NJ および MP 法による解析は long branch attraction の影響を強く受けてしまったと言える．

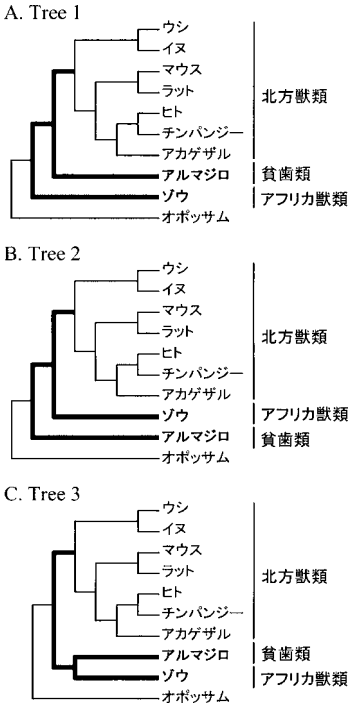

図 2. 真獣類の初期進化に関する 3 つの系統仮説．北方獣類(ウシ，イヌ，マウス，ラット，ヒト，チンパンジー，アカゲザル)内部の系統関係は固定した．

表 1. 三つの系統仮説に対する各モデルによる対数尤度の比較．系統仮説の番号は図 2 に従う．Concatenate model では最尤系統樹が置換モデルによって異なるが(A)，Separate model (2,789 遺伝子間で異なるモデル) では一貫して Tree 1 を支持した (B)．最尤系統樹における対数尤度は括弧内に示し，他の系統樹の対数尤度の差および 1SE 値は Kishino-Hasegawa test により推定した．KH と wSH は，CONSEL を用いて計算した Kishino-Hasegawa test および weighted test of Shimodaira-Hasegawa による P 値を示す．K は各モデルのパラメータ数を示し，AIC は赤池情報量規準を示す．

Concatenate or Separate model	置換モデル	Tree	$< \ln L > (\Delta \ln L \pm SE)$	KH	wSH	BP	K	AIC
Concatenate model	GTR + Γ_8	1	−117.2 ± 31.1	0.000	0.000	0.0		
		2	−147.3 ± 29.7	0.000	0.000	0.0		
		3	<−4,076,316.3>			100.0	26	8,152,684.6
	Codon + Γ_4	1	<−3,828,351.7>			88.1	81	7,656,865.4
		2	−77.8 ± 64.5	0.112	0.185	11.3		
		3	−142.7 ± 65.0	0.014	0.026	0.6		
	JTT-F + Γ_8	1	<−1,905,933.9>			51.6	37	3,811,941.8
		2	−84.1 ± 37.4	0.014	0.028	0.2		
		3	−1.7 ± 41.9	0.478	0.637	48.2		
Separate model (2,789遺伝子)	GTR + Γ_8	1	<−3,963,489.9>			86.2	72,514	8,072,007.8
		2	−117.4 ± 72.3	0.050	0.092	4.1		
		3	−91.4 ± 72.7	0.104	0.174	9.7		
	Codon + Γ_4	1	<−3,621,322.1>			89.6	225,909	7,694,462.2
		2	−128.0 ± 103.2	0.107	0.164	10.4		
		3	−527.9 ± 96.3	0.000	0.000	0.0		
	JTT-F + Γ_8	1	<−1,799,245.4>			93.4	103,193	3,804,876.8
		2	−134.9 ± 88.5	0.064	0.112	6.6		
		3	−317.6 ± 85.5	0.000	0.000	0.0		

5. 全配列を結合したデータセットを用いた最尤法による系統解析

次に 2,789 遺伝子を結合して 1 つの配列とし (Concatenate model)，最尤法による系統解析をおこなった．本解析に用いるモデルとしては，塩基置換モデル (GTR+Γ_8)，コドン置換モデル (+Γ_4) (Yang et al., 1998)，アミノ酸置換モデル (JTT-F+Γ_8) を採用した．なお，最尤法においては図 2 に示した 3 つの系統仮説に関して解析し，北方獣類内の関係は固定した．以降，アフリカ獣類が最初に分岐したとする仮説を Tree 1，貧歯類が最初に分岐したとする仮説を Tree 2，北方獣類が最初に分岐したとする仮説を Tree 3 と呼ぶ．PAML3.15 (Yang, 1997) を用いて各モデルに基づく計算をおこなった結果，非常に面白いことに，解析方法ごとに全く異なる系統樹が支持された (表 1 および図 3)．まず塩基配列を Concatenate model で解析する手法は一般的によく用いられるが，この解析方法では，Tree 3 (アフリカ獣類と貧歯類が単系統) が非常に強く支持された (BP= 100%)．同時に他の 2 仮説は weighted test of Shimodaira and Hasegawa (wSH) (Shimodaira and Hasegawa, 1999) でも強く棄却された ($P < 0.001$，BP= 0.0%)．この解析モデルは一般的によく用いられる方法であるため，もしこの解析のみをおこなっていたら Tree 3 が正しいであろうと結論付けてしまっていたかもしれない．しかしながら，意外なことにコドン置換モデルでは Tree 3 が棄却されてしまい (BP= 0.6%，P= 0.026 (wSH test))，代わりに Tree 1 が最尤系統樹となった．一方でアミノ酸置換モデルでは，Tree 2 が棄却され (BP= 0.2%)，他の 2 仮説はほぼ同程度に支持された．このように，本研究で用いる 2,789 遺伝子 (約 1 Mbp) を結合させたデータセットは，仮定する置換モデルの影響を強く受けやすいことが示された．

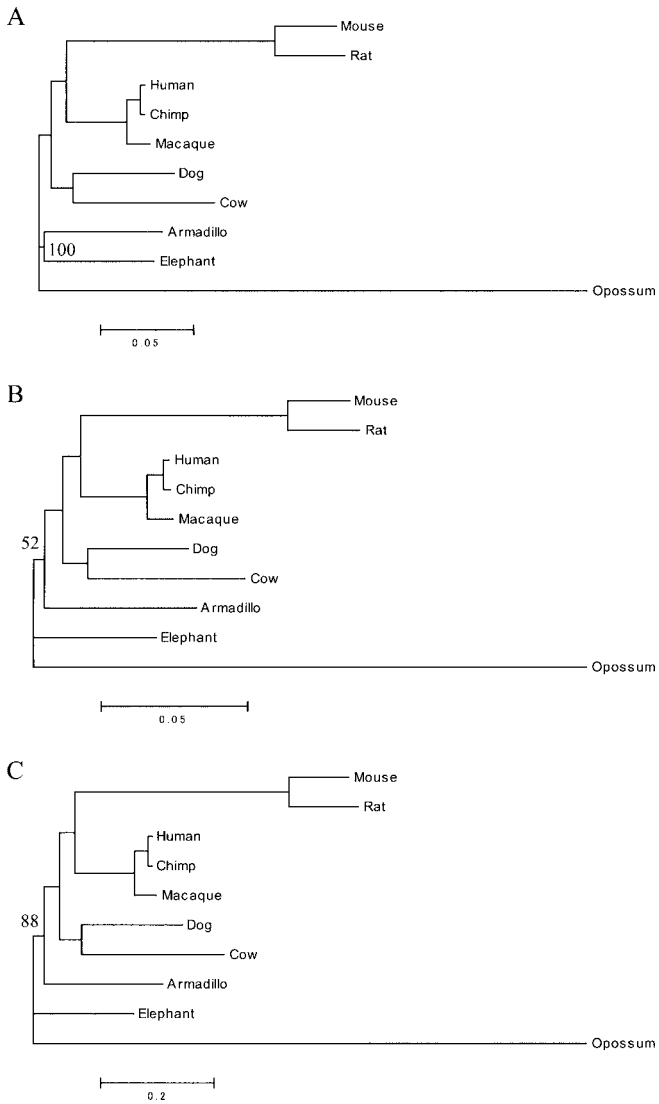

図 3. Concatenate model による最尤系統樹. (A) 塩基置換モデル (GTR+Γ_8), (B) アミノ酸置換モデル (JTT-F+Γ_8), (C) コドン置換モデル (+Γ_4) (図 1 参照).

6. 遺伝子データセットを分割した解析

本研究で収集したデータセットは非常に多くの遺伝子から構成されており, 各遺伝子間の進化速度や進化パターンの違いは非常に大きいと予想される. そこで, 次に 2,789 個の遺伝子ごとに異なるパラメータを与えて計算する Separate model を用いて最尤解析をおこなった (Kishino and Hasegawa, 1989). ここでは MOLPHY (Adachi and Hasegawa, 1996) に含まれている TotalML プログラムを用いて全データセットの対数尤度を計算した. また Kishino-Hasegawa test (KH) (Kishino and Hasegawa, 1989) および weighted test of Shimodaira-Hasegawa (wSH) (Shimodaira

表 2. Concatenate および Separate model で解析した際の各系統仮説に対する BP 値の比較．#c，K，n はそれぞれ Separate model における遺伝子カテゴリ数（#c =1 の場合は Concatenate model），パラメータ数，形質数（サイト数）を表す．この遺伝子カテゴリは 2,789 遺伝子それぞれの全枝長に基づいてグループ分けをおこなった．n/K が 40 より大きい場合は AIC を，40 に満たない場合は AICc を斜体で表示した．また，AIC もしくは AICc が最小になる行を太字で表示した．

A. 塩基置換モデル (GTR + Γ_8)

#c	Ln L	K	n	n/K	AIC	AICc	Tree 1	Tree 2	Tree 3
1	-4076316.3	26	1011870	38918.1	*8152684.6*	8152684.6	0.0	0.0	100.0
5	-4059904.9	130	1011870	7783.6	*8120069.8*	8120069.8	0.0	0.0	100.0
10	-4058547.6	260	1011870	3891.8	*8117615.2*	8117615.3	0.0	0.0	100.0
56	-4055469.5	1456	1011870	695.0	*8113851.0*	8113855.2	0.1	0.0	99.9
100	-4053634.1	2600	1011870	389.2	*8112468.2*	8112481.6	0.1	0.0	99.9
200	-4049237.9	5200	1011870	194.6	*8108875.8*	8108929.5	0.2	0.0	99.8
558	-4035535.0	14508	1011870	69.7	*8100086.0*	8100508.1	1.7	0.0	98.3
930	-4022303.0	24180	1011870	41.8	*8092966.0*	8094150.0	3.6	0.0	96.4
1395	-4006623.4	36270	1011870	27.9	8085786.8	*8088483.7*	25.0	0.7	74.3
2789	**-3963489.9**	**72514**	**1011870**	**14.0**	**8072007.8**	***8083203.5***	**86.2**	**4.1**	**9.7**

B. コドン置換モデル (+ Γ_4)

#c	Ln L	K	n	n/K	AIC	AICc	Tree 1	Tree 2	Tree 3
1	-3828351.7	81	337290	4164.1	*7656865.4*	7656865.4	88.1	11.3	0.6
5	-3810589.3	405	337290	832.8	*7621988.6*	7621989.6	94.3	5.1	0.7
10	-3808198.7	810	337290	416.4	*7618017.4*	7618021.3	93.3	5.9	0.8
56	-3802941.9	4536	337290	74.4	*7614955.8*	7615079.5	93.0	5.2	1.7
100	**-3799324.6**	**8100**	**337290**	**41.6**	***7614849.2***	**7615247.9**	**94**.0	**4.9**	**1.1**
200	-3791928.7	16200	337290	20.8	7616257.4	*7617892.2*	91.0	8.1	1.0
558	-3766336.0	45198	337290	7.5	7623068.0	*7637056.1*	96.7	2.9	0.3
930	-3741173.9	75330	337290	4.5	7633007.8	*7676332.8*	98.0	1.7	0.3
1395	-3712084.5	112995	337290	3.0	7650159.0	*7764009.4*	96.2	3.8	0.0
2789	-3621322.1	225909	337290	1.5	7694462.2	*8610876.3*	89.6	10.4	0.0

C. アミノ酸置換モデル (JTT-F + Γ_8)

#c	Ln L	K	n	n/K	AIC	AICc	Tree 1	Tree 2	Tree 3
1	-1905933.9	37	337290	9115.9	*3811941.8*	3811941.8	51.6	0.2	48.2
5	-1879320.4	185	337290	1823.2	*3759010.8*	3759011.0	63.4	0.2	36.5
10	-1877405.7	370	337290	911.6	*3755551.4*	3755552.2	63.9	0.3	35.9
56	**-1875094.5**	**2072**	**337290**	**162.8**	***3754333.0***	**3754358.6**	**56.6**	**0.1**	**43.2**
100	-1873607.4	3700	337290	91.2	*3754614.8*	3754696.9	58.7	0.5	40.9
200	-1870213.5	7400	337290	45.6	*3755227.0*	3755559.0	59.8	0.2	40.1
558	-1858842.6	20646	337290	16.3	3758977.2	*3761669.7*	81.2	1.1	17.7
930	-1847528.8	34410	337290	9.8	3763877.6	*3771696.4*	81.6	6.5	11.9
1395	-1834624.0	51615	337290	6.5	3772478.0	*3791129.7*	87.1	10.9	2.0
2789	-1799245.4	103193	337290	3.3	3804876.8	*3895855.7*	93.4	6.6	0.0

and Hasegawa, 1999）は，CONSEL（Shimodaira and Hasegawa, 2001）を用いて計算した．ブートストラップ値（BP）は RELL 法（Kishino et al., 1990）を用いて 1 万回のリサンプリングをおこなった．なお，モデル選択の規準としては赤池情報量規準（AIC）を採用した（Akaike, 1973）．

$$AIC = -2\log L + 2K \qquad (L = likelihood,\ K = \text{パラメータ数})$$

その結果，面白いことに塩基，アミノ酸，コドン置換モデルのいずれにおいても Tree 1 が支

持された(表 1)．また塩基およびアミノ酸置換モデルでは遺伝子ごとに分割した解析は AIC に基づくとより良いモデルであることが示された．コドン置換モデルではおそらくパラメータ数が多すぎたことが原因で AIC 値の減少は見られなかったが，分割するかしないかに関わらず Tree 1 を支持するという結果となった．

コドン置換モデルでも見られたように，2,789 個の遺伝子それぞれに分割するとパラメータ数がかなり増大するため，その分割が AIC の規準から良いモデルであるとは言い切れない．そこで，この 2,789 遺伝子をその進化速度ごとにいくつかのカテゴリに分け，そのカテゴリごとに異なるパラメータを与えた解析をおこなった．すなわち，まず遺伝子ごとの ML 解析で得られる全枝長(total branch length)を指標として，その大きいものからカテゴリ分けをおこなった．その際，5，10，56，100，200，558，930，1395，2,789 個のカテゴリに分割し，それぞれに関して Separate model を適用した．またこの解析の際，AIC ではパラメータ数が非常に多いモデルが良いモデルとみなされる傾向があるため，その偏りを補正した AICc も採用した．系統解析の場合，パラメータ数(K)に対してサイト数(n)が比較的大きければ($n/K > 40$) AIC を採用し，小さければ($n/K < 40$) AICc を使うことが推奨されている(Burnham and Anderson, 2003; Posada and Buckley, 2004)．

$$AICc = AIC + \frac{2K(K+1)}{n-K-1} \qquad (K = \text{パラメータ数}, n = \text{形質数(サイト数)})$$

表 2 は，尤度および AIC (もしくは AICc)を比較した結果である．塩基置換モデルでは，結局 AICc を規準とした場合でも 2,789 遺伝子ごとに異なるパラメータを与えたモデルが最適であるという結果が得られ，Tree 1 が支持された(BP=86%)．コドン置換モデルでは n/K の値から AIC を規準として用い，100 カテゴリの分割モデルが最適であると示されたが，これも Tree 1 を支持した(BP=94%)．またアミノ酸レベルでは 56 カテゴリに分割した場合に Tree 1 が最尤系統樹であったが，Tree 3 とほぼ同程度に支持される結果になった．したがって，遺伝子をカテゴリ分けした場合においても AIC または AICc に基づいて Tree 1 が最も可能性が高いことが示された．

7. 進化速度の速い遺伝子を除去した場合

進化速度の速い遺伝子は，long branch attraction，塩基組成の偏り，heterotachy などの影響を大きくし，それが原因で誤った系統樹を支持してしまう場合がしばしばある(Delsuc et al., 2006; Philippe et al., 2005; Brinkmann et al., 2005)．したがって，そうした遺伝子を除いて Concatenate model を適用した場合，Separate model と同じ仮説を支持するかもしれない(Philippe et al., 2005; Brinkmann et al., 2005)．そこで，全遺伝子の中から進化速度の速い遺伝子を除いた場合に各系統仮説に対する支持がどう変化するかを調べた．具体的には，2,789 遺伝子データセットの中から進化速度の速い順(全枝長の大きい順)に 50 遺伝子ずつ除いて 56 個のデータセットを作り，それぞれに関して Concatenate model を用いた最尤解析をおこなって各系統仮説に対する BP 値の推移を調べた．その結果，塩基レベルで非常に強く支持されていた Tree 3 の BP 値が進化速度の速い遺伝子を除くとゼロにまで急激に下がり，代わりに Tree 1 や Tree 2 (特に Tree 1)に対する BP 値が増大していった(図 4 (a))．また，アミノ酸では Tree 1 や 3 に対してほぼ同程度の支持を示していたが，進化速度の速い遺伝子を除くと Tree 3 の BP 値が急激に下がり，代わりに Tree 1 の BP 値が増大した(図 4 (c))．さらにコドン置換モデルにおいてのみ計算時間短縮のために 100 遺伝子ずつ除いて 28 個のデータセットに関して解析したが，結局 Tree 3 が棄却されるという結果に変化はなかった(図 4 (b))．なお，いずれの解析でも遺伝子を大量に除くと各系統仮説に対する支持が曖昧になるが，これは進化速度の遅い

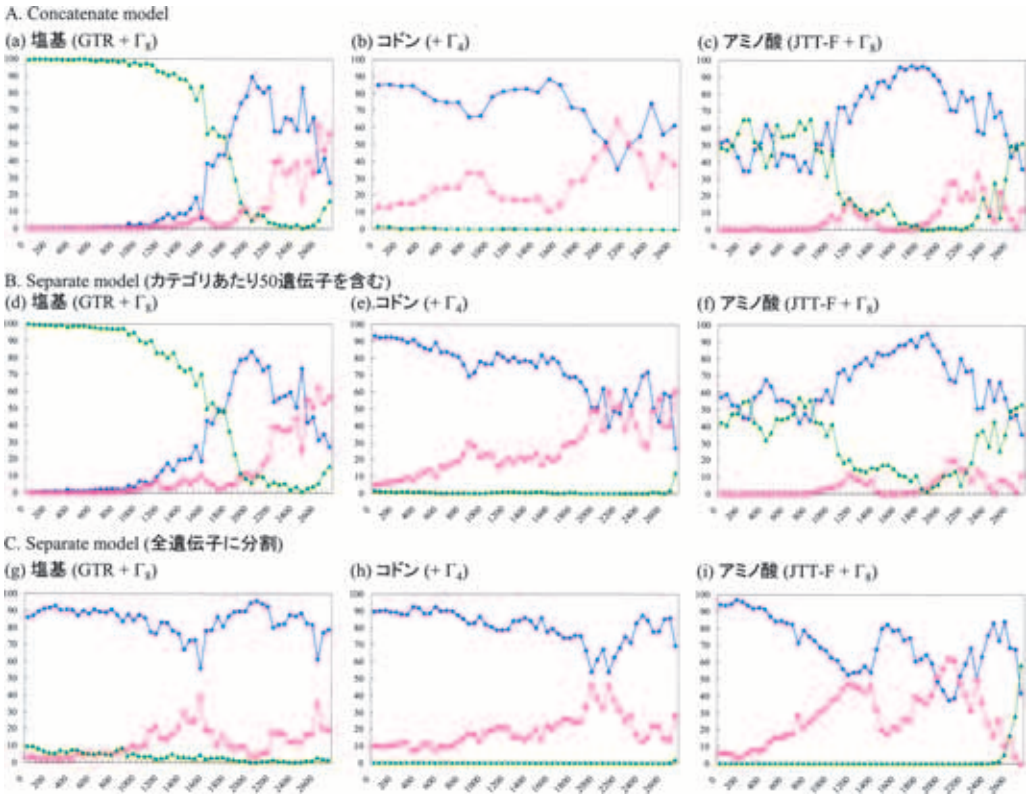

図 4. 全枝長の大きな遺伝子を 50 ずつ除去して解析した際の各系統仮説に対する BP 値の推移．横軸は全 2,789 遺伝子データセットの中から除いた遺伝子数を示す．各色は，Tree 1 (青)，Tree 2 (ピンク)，Tree 3 (黄緑) に対する値を示す．(A) Concatenate model, (B) データセットを全枝長ごとに 50 遺伝子ずつのカテゴリに分け Separate model で解析，(C) データセットをそれぞれの遺伝子ごとに Separate model で解析．各データセットに関して，塩基置換モデル (GTR+Γ_8; a, d, g)，コドン置換モデル (+Γ_4; b, e, h)，アミノ酸置換モデル (JTT+Γ_8; c, f, i) で解析した．

遺伝子群において系統関係を示すサイトが少なすぎることが原因であると考えられる．

さらにこの 56 個のデータセットに関して，進化速度ごとに 50 個ずつの遺伝子をカテゴリとして分け，各カテゴリに異なるパラメータを与える Separate model を適用した．その結果，BP 値の推移は Concatenate model の場合と非常に類似していた．特にアミノ酸置換モデルでは，全遺伝子を用いた場合は 56 カテゴリの分割が AIC において最適モデルであり Tree 1 と Tree 3 を同程度に支持する結果であったが，進化速度の速い遺伝子を除くと Tree 1 に対する BP が増大した (図 4 (f))．さらに，この 56 個のデータセットに関して，それぞれの遺伝子ごとに分割した Separate model も適用した．塩基配列レベルではこの全遺伝子に分割した解析が AICc の規準で最適モデルとなる．塩基，コドン，アミノ酸のいずれの解析においても，進化速度の速い遺伝子を除いた場合は Tree 3 が支持されなかった (図 4 (g–i))．

以上のように，本研究で用いた膨大な遺伝子データセットはモデルごとに大きく異なる結果を示すことが示された．塩基配列の Concatenate model は一般的によく用いられる解析法であるが，それによると Tree 3 が非常に強く支持された (BP=100%)．それに対して，遺伝子ごと，あるいは遺伝子カテゴリごとに分割して別々のパラメータを与える Separate model では，

AIC および AICc の規準により Tree 1 を支持するモデルが最適モデルとなった．この結果から，Tree 2 や 3 の可能性を棄却できたわけではないが，Tree 1（アフリカ獣類が最初に分岐）の可能性が最も高いことが示された．本研究と類似して，Hallstrom et al.（2007）は最近 2,840 遺伝子データを用いて真獣類の初期進化の問題に取り組み，Concatenate model を用いた解析によりアフリカ獣類と貧歯類が単系統であると結論付けている．しかしながら本研究の解析結果を考慮すると，彼らのデータも Separate model を適用するなどして詳細に解析すれば異なる結果が得られる可能性があると考えられる．

8. 結果の違いを引き起こした原因は何か？

全ゲノム規模のデータを用いた場合でも，誤った系統樹を支持してしまう要因がいくつか考えられる．(1)塩基もしくはアミノ酸組成の偏り（Rokas et al., 2003; Phillips et al., 2004; Jeffroy et al., 2006），(2)系統ごとに進化速度が大きく異なる場合に起こる long branch attraction（Soltis et al., 2004; Dopazo and Dopazo, 2005; Philippe et al., 2005; Felsenstein, 1978），(3)解析に用いる種のサンプリングの偏り（Soltis et al., 2004; Blair et al., 2002; Philippe et al., 2005），(4) Heterotachy（サイトごとの進化速度の変化）（Lopez et al., 2002; Philippe et al., 2005; Brinkmann et al., 2005; Spencer et al., 2005; Kolaczkowski and Thornton, 2004; Lockhart et al., 2006; Shalchian-Tabrizi

表 3. TREE-PUZZLE を用いた，アミノ酸・塩基組成を平均値と比較した χ^2 検定.

(A) アミノ酸

Species	*P*-value of χ^2 test
Human	97.59%
Chimpanzee	97.35%
Macaque	97.56%
Mouse	0.00%
Rat	0.00%
Dog	59.92%
Cow	0.02%
Armadillo	91.13%
Elephant	99.99%
Opossum	0.00%

(B) 第一コドン座位

Species	*P*-value of χ^2 test
Human	39.93%
Chimpanzee	34.89%
Macaque	43.06%
Mouse	0.13%
Rat	0.03%
Dog	0.16%
Cow	0.00%
Armadillo	30.99%
Elephant	99.78%
Opossum	0.00%

(C) 第二コドン座位

Species	*P*-value of χ^2 test
Human	28.59%
Chimpanzee	29.27%
Macaque	23.16%
Mouse	0.11%
Rat	0.01%
Dog	29.54%
Cow	0.25%
Armadillo	92.38%
Elephant	82.46%
Opossum	0.00%

(D) 第三コドン座位

Species	*P*-value of χ^2 test
Human	0.00%
Chimpanzee	0.00%
Macaque	0.00%
Mouse	0.00%
Rat	0.00%
Dog	0.00%
Cow	0.00%
Armadillo	0.00%
Elephant	44.69%
Opossum	0.00%

表 4. げっ歯類またはウシを除き Concatenate model (GTR+Γ_8) を用いた解析の対数尤度の比較.

(A) ウシを除外

Tree	$< \ln L > (\Delta \ln L \pm SE)$	KH	wSH	BP (%)
1	−147.6 ± 33.3	0.000	0.000	0.0
2	−178.6 ± 32.0	0.000	0.000	0.0
3	< −3,779,962.6 >			100.0

(B) げっ歯類（マウス・ラット）を除外

Tree	$< \ln L > (\Delta \ln L \pm SE)$	KH	wSH	BP (%)
1	−86.6 ± 28.0	0.001	0.002	0.1
2	−120.9 ± 26.3	0.000	0.000	0.0
3	< −3430106.9>			99.9

(C) げっ歯類およびウシを除外

Tree	$< \ln L > (\Delta \ln L \pm SE)$	KH	wSH	BP (%)
1	−125.2 ± 30.3	0.000	0.000	0.0
2	−159.8 ± 28.7	0.000	0.000	0.0
3	< −3127148.1>			100.0

et al., 2006) である．もし本解析において long branch attraction が作用しているならば，系統樹上で枝長の違いが明確に現れるはずである．本研究のデータセットにおいて塩基配列を Concatenate model で解析した際の系統樹は図 3 であるが，ここでは北方獣類内のマウス，ラット，ウシにおいて長い枝長が見られた．一方でアミノ酸や塩基組成の偏りを TREE-PUZZLE (Strimmer and von Haeseler, 1996) を用いて解析すると，同じくマウス，ラット，ウシにおいて大きな偏りが見られる (表 3)．そこでげっ歯類もしくはウシの配列を除き，塩基レベルで Concatenate model を用いた解析をおこなった．もし Tree 3 に対する強い支持の原因がこれらの long branch attraction やアミノ酸・塩基組成の偏りであれば，これらを除いた場合には Tree 1 や 2 が支持され，Tree 3 に対する支持は減少すると期待される．しかしながら，結果的には表 4 に示すようにこれらを除いても Tree 3 が相変わらず非常に強く支持された．したがって，Tree 3 を強く支持するという結果は，long branch attraction やアミノ酸・塩基組成の偏りが原因ではないだろうと考えられる．しかも，仮にそうした要因が悪影響を及ぼしているならば，Separate model によってこの誤りが改善されることは考えにくい．こうしたことから，本研究で示した結果の違いは，種間ではなく遺伝子間の差異に原因があると考えられる．

もう一つの可能性としては，パラログ遺伝子がデータセット内に紛れ込んでいることが考えられる．もしそれが影響しているならば，Tree 3 を支持する遺伝子群には特にパラログ遺伝子データが多く混入していると期待され，そのためその遺伝子の全枝長は他の遺伝子群よりも比較的大きくなるだろうと予想される．そこで，Tree 3 を支持する 848 遺伝子それぞれの全枝長の分布を全 2,789 遺伝子のものと比較した．この際の全枝長は PAML3.15 を用いて GTR+Γ_8 モデルで解析したものを指標としている．結果的には図 5 に示すように遺伝子の全枝長の分布に明確な違いは見られず，パラログ遺伝子が Tree 3 を支持する遺伝子に特に多いという結論は見られなかった．したがって，これまでの解析からは Concatenate model において Tree 3 を支持した原因を明確に結論づけることはできなかった．

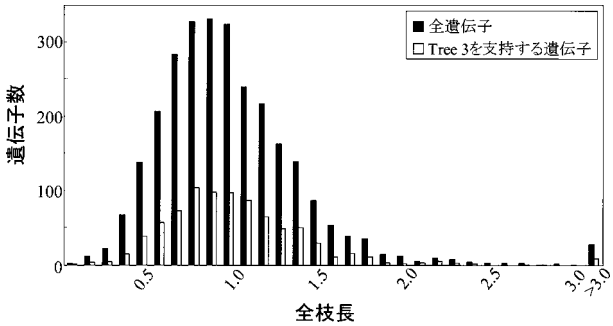

図 5. Tree 3 を支持する 848 遺伝子(白)と全 2,789 遺伝子(黒)の全枝長の分布．各遺伝子の全枝長は PAML 3.15 を用いて GTR+Γ_8 モデルで解析した．

9. 総括

全ゲノム規模の配列データが蓄積されるにつれて系統解析に用いられる遺伝子数も膨大になってきており，それにつれて遺伝子間の進化速度・進化パターンの差が系統樹推定に及ぼす影響と解決策に関して注目されてきている(Gadagkar et al., 2005; Seo et al., 2005)．本研究ではそうした膨大量の遺伝子配列を結合したデータセットを用いた場合に置換モデルによって結果が大きく異なるという極端な例を示した．またその原因は long branch attraction や配列組成の偏りではなく，おそらく遺伝子ごとの進化速度・進化パターンの大きな違いであると考えられる．系統解析におけるこうしたエラーは long branch attraction や配列組成の偏りよりも検出しにくいが，おそらく一般的な現象であると考えられる．さらにこうしたエラーは遺伝子ごとに異なるパラメータを仮定する Separate model を適用すれば改善できることも示された．以上のように，系統解析において膨大な配列情報は非常に有用であり標本誤差を小さくすることができるものの，単純に配列データを結合して解析してしまうと重大な誤りを引き起こす危険性がある．したがって系統関係を明らかにするためにはデータ解析に十分に注意を払い，遺伝子ごとの進化速度・進化パターンの違いを考慮に入れた Separate model を用いるなどしてモデルを改善していくことが必要である．

前述したように，真獣類の初期の系統関係を解明することは大規模な大陸移動との関連性を明らかにする上で非常に重要である．本研究では Tree 1 が最も可能性が高いことが示されたが，他の 2 つの仮説を完全に否定できるものではなかった．こうした全ゲノム規模のデータと適用できる最適なモデルを用いても結論付けられないことを考えると，今後はさらに多くの種を系統解析に用い，より多くの遺伝子データをさらに良い進化モデルを用いて解析することが必要となる．実際，数十種の哺乳類においてゲノムプロジェクトが現在進行中であるため，より多くの種を用いることが当面可能な改善策であると考えられる．またこうした全ゲノムデータが蓄積されれば遺伝子データのみならず，レトロポゾンの挿入解析のような手法(Nikaido et al., 1999; Shedlock and Okada, 2000; Nikaido et al., 2001; Nishihara et al., 2005; Kriegs et al., 2006; Sasaki et al., 2006; Nishihara et al., 2006; Murphy et al., 2007)も含めて多角的にアプローチすることが可能となる．こうしたゲノムデータを有効に解析することで，哺乳類の進化の過程，特に真獣類の分岐と大陸移動との関連性も近い将来明らかになることと期待される．

謝辞

本研究の一部は，文部科学省科学研究費補助金，および日本学術振興会ならびに情報・システム研究機構新領域融合研究センターの研究費を用いておこなわれた．

参 考 文 献

Adachi, J. and Hasegawa, M. (1996). MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood, *Computer Science Monographs*, No. 28, 1–150.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), 267–281, Akademiai Kiado, Budapest.

Amrine-Madsen, H., Koepfli, K. P., Wayne, R. K. and Springer, M. S. (2003). A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships, *Molecular Phylogenetics and Evolution*, **28**, 225–240.

Blair, J. E., Ikeo, K., Gojobori, T. and Hedges, S. B. (2002). The evolutionary position of nematodes, *BMC Evolutionary Biology*, **2**, 7.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D. and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner, *Genome Research*, **14**, 708–715.

Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G. and Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics, *Systematic Biology*, **54**, 743–757.

Burnham, K. P. and Anderson, D. R. (2003). *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach*, Springer, New York.

Cao, Y., Fujiwara, M., Nikaido, M., Okada, N. and Hasegawa, M. (2000). Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data, *Gene*, **259**, 149–158.

Delsuc, F., Scally, M., Madsen, O., Stanhope, M. J., de Jong, W. W., Catzeflis, F. M., Springer, M. S. and Douzery, E. J. (2002). Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting, *Molecular Biology and Evolution*, **19**, 1656–1671.

Delsuc, F., Brinkmann, H. and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life, *Nature Reviews Genetics*, **6**, 361–375.

Delsuc, F., Brinkmann, H., Chourrout, D. and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates, *Nature*, **439**, 965–968.

Dopazo, H. and Dopazo, J. (2005). Genome-scale evidence of the nematode-arthropod clade, *Genome Biology*, **6**, R41.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading, *Systematic Zoology*, **27**, 401–410.

Gadagkar, S. R., Rosenberg, M. S. and Kumar, S. (2005). Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree, *Journal of Experimental Zoology Part B*: *Molecular and Developmental Evolution*, **304**, 64–74.

Hallstrom, B., Kullberg, M., Nilsson, M. and Janke, A. (2007). Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups, *Molecular Biology and Evolution*, **24**, 2059–2068.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D. and Kent, W. J. (2006). The UCSC genome browser database: Update 2006, *Nucleic Acids Research*, **34**, D590–598.

Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H. (2006). Phylogenomics: The beginning of incongruence?, *Trends in Genetics*, **22**, 225–231.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea, *Journal of Molecular Evolution*, **29**, 170–179.

Kishino, H., Miyata, T. and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *Journal of Molecular Evolution*, **31**, 151–160.

Kolaczkowski, B. and Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous, *Nature*, **431**, 980–984.

Kriegs, J. O., Churakov, G., Kiefmann, M., Jordan, U., Brosius, J. and Schmitz, J. (2006). Retroposed elements as archives for the evolutionary history of placental mammals, *PLoS Biology*, **4**, e91.

Kumar, S. and Hedges, S. B. (1998). A molecular timescale for vertebrate evolution, *Nature*, **392**, 917–920.

Lockhart, P., Novis, P., Milligan, B. G., Riden, J., Rambaut, A. and Larkum, T. (2006). Heterotachy and tree building: A case study with plastids and eubacteria, *Molecular Biology and Evolution*, **23**, 40–45.

Lopez, P., Casane, D. and Philippe, H. (2002). Heterotachy, an important process of protein evolution, *Molecular Biology and Evolution*, **19**, 1–7.

Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W. and Springer, M. S. (2001). Parallel adaptive radiations in two major clades of placental mammals, *Nature*, **409**, 610–614.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. and O'Brien, S. J. (2001a). Molecular phylogenetics and the origins of placental mammals, *Nature*, **409**, 614–618.

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W. and Springer, M. S. (2001b). Resolution of the early placental mammal radiation using Bayesian phylogenetics, *Science*, **294**, 2348–2351.

Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S. and Miller, W. (2007). Using genomic data to unravel the root of the placental mammal phylogeny, *Genome Research*, **17**, 413–421.

Nikaido, M., Rooney, A. P. and Okada, N. (1999). Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: Hippopotamuses are the closest extant relatives of whales, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 10261–10266.

Nikaido, M., Matsuno, F., Hamilton, H., Brownell, R. L., Jr., Cao, Y., Ding, W., Zuoyan, Z., Shedlock, A. M., Fordyce, R. E., Hasegawa, M. and Okada, N. (2001). Retroposon analysis of major cetacean lineages: The monophyly of toothed whales and the paraphyly of river dolphins, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 7384–7389.

Nishihara, H., Satta, Y., Nikaido, M., Thewissen, J. G., Stanhope, M. J. and Okada, N. (2005). A retroposon analysis of Afrotherian phylogeny, *Molecular Biology and Evolution*, **22**, 1823-1833.

Nishihara, H., Hasegawa, M. and Okada, N. (2006). Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 9929–9934.

Philippe, H., Lartillot, N. and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia, *Molecular Biology and Evolution*, **22**, 1246–1253.

Phillips, M. J., Delsuc, F. and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases, *Molecular Biology and Evolution*, **21**, 1455–1458.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests, *Systematic Biology*, **53**, 793–808.

Rokas, A., Williams, B. L., King, N. and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies, *Nature*, **425**, 798–804.

Sasaki, T., Yasukawa, Y., Takahashi, K., Miura, S., Shedlock, A. M. and Okada, N. (2006). Extensive morphological convergence and rapid radiation in the evolutionary history of the family Geoemydidae (old world pond turtles) revealed by SINE insertion analysis, *Systematic Biology*, **55**, 912–927.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W. (2003). Human-mouse alignments with BLASTZ, *Genome Research*, **13**, 103–107.

Seo, T.-K., Kishino, H. and Thorne, J. L. (2005). Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4436–4441.

Shalchian-Tabrizi, K., Skanseng, M., Ronquist, F., Klaveness, D., Bachvaroff, T. R., Delwiche, C. F., Botnen, A., Tengs, T. and Jakobsen, K. S. (2006). Heterotachy processes in rhodophyte-derived secondhand plastid genes: Implications for addressing the origin and evolution of dinoflagellate plastids, *Molecular Biology and Evolution*, **23**, 1504–1515.

Shedlock, A. M. and Okada, N. (2000). SINE insertions: Powerful tools for molecular systematics, *Bioessays*, **22**, 148–160.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Molecular Biology and Evolution*, **16**, 1114–1116.

Shimodaira, H. and Hasegawa, M. (2001). CONSEL: For assessing the confidence of phylogenetic tree selection, *Bioinformatics*, **17**, 1246–1247.

Smith, A. G., Smith, D. G. and Funnell, B. M. (2004). *Atlas of Cenozoic and Mesozoic Coastlines*, Cambridge University Press, New York.

Soltis, D. E., Albert, V. A., Savolainen, V., Hilu, K., Qiu, Y. L., Chase, M. W., Farris, J. S., Stefanovic, S., Rice, D. W., Palmer, J. D. and Soltis, P. S. (2004). Genome-scale data, angiosperm relationships, and "ending incongruence": A cautionary tale in phylogenetics, *Trends in Plant Science*, **9**, 477–483.

Spencer, M., Susko, E. and Roger, A. J. (2005). Likelihood, parsimony, and heterogeneous evolution, *Molecular Biology and Evolution*, **22**, 1161–1164.

Springer, M. S., Stanhope, M. J., Madsen, O. and de Jong, W. W. (2004). Molecules consolidate the placental mammal tree, *Trends in Ecology & Evolution*, **19**, 430–438.

Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies, *Molecular Biology and Evolution*, **13**, 964–969.

Waddell, P. J., Okada, N. and Hasegawa, M. (1999). Towards resolving the interordinal relationships of placental mammals, *Systematic Biology*, **48**, 1–5.

Waddell, P. J. and Shelley, S. (2003). Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models, *Molecular Phylogenetics and Evolution*, **28**, 197–224.

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood, *Computer Applications in the Biosciences*, **13**, 555–556.

Yang, Z., Nielsen, R. and Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution, *Molecular Biology and Evolution*, **15**, 1600–1611.

Proceedings of the Institute of Statistical Mathematics Vol. 56, No. 1, 19–35 (2008) 35

Power and Pitfalls of Phylogenomics: Lessons from a Genome-scale Analysis with Respect to the Root of the Eutherian Tree

Hidenori Nishihara[1], Norihiro Okada[1] and Masami Hasegawa[2]

[1]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology
[2]School of Life Sciences, Fudan University

In the post-genomic era, genome-scale approaches to phylogenetic inference (phylogenomics) are being applied extensively to overcome sampling errors. Sampling error vanishes as the number of genes provided for the analysis increases, but the fully resolved tree can still be wrong if the phylogenetic inference is biased (systematic error). In the present study, we collected 2,789 genes (1 Mbp) from 10 mammalian genomic sequences by screening whole-genome data, and performed an extensive maximum likelihood (ML) analysis to determine the root of the eutherian tree. The conventional method of concatenate analysis of nucleotide sequences strongly suggests a misled monophyly of Afrotheria (e.g., elephant) and Xenarthra (e.g., armadillo). However, this tree is not supported by a "Separate model" that takes into account the different tempos and modes of evolution among genes, and instead the basal Afrotheria tree is favored. This analysis demonstrates that the separate model, rather than the concatenate model, should be used in cases of phylogenetic inference for genome-scale data.

Key words: Phylogenomics, maximum likelihood, separate model, mammalian phylogeny.