統計数理（2002）
第 50 巻 第 1 号 17–31

特集「DNA 配列の統計解析」
［研究詳解］

分子進化速度のベイズ型階層モデル

岸野 洋久[1]・Jeffrey L. Thorne[2]

（受付 2001 年 11 月 29 日；改訂 2002 年 1 月 28 日）

要　　旨

進化の過程で生物は，進化速度とその変動の形で多様化と適応の痕跡をゲノムに残す．本稿ではまず，トウモロコシの栽培化における選択圧，ハワイにおける silversword 群団の適応放散と調節遺伝子の加速化，ウイルスの免疫適応過程，遺伝子重複の運命など，最近の研究を簡単に紹介しながら，進化研究における速度変動の推定の持つ役割を確認する．続いて筆者らの提案した進化速度の確率変動を記述する階層モデルを紹介し，その性能を評価する．最後に，ゲノムデータベース解析における階層モデルの可能性を検討する．

キーワード：分子進化速度の確率変動，階層モデル，マルコフ連鎖モンテカルロ法，
共進化の検出，複数遺伝子モデル，ゲノムデータベース解析．

1. はじめに：速度変化に見る生物の適応・多様化の痕跡

本稿では著者らが開発した分子進化の階層モデルとその性能評価（Thorne et al.（1998），Kishino et al.（2001），Thorne and Kishino（2002））を中心に概説するが，生物の適応進化と多様化がゲノムに残した痕跡を理解する上で分子進化速度の変動を調べることが有効であることを論述することに力点を置く．ゲノムデータベース解析において浮き彫りになってくるであろう統計的問題に言及する．

Kimura（1983）の「分子進化の中立説」は分子レベルでの進化についての理解の変革をもたらした．中立説は，塩基置換や挿入・欠失，遺伝子領域の転座や逆位など，ゲノムを変化させる突然変異のうち，既存のゲノムに比して優勢のものは稀で，その多くのものが有害か同等であるとする．表現型や機能レベルでは生存競争による淘汰が進化の主たる原動力になっているのに対して，たんぱく質の種間比較や多型性の分布など，数々のデータが大まかにおいて分子進化の中立説を支持している．中立説の下では，突然変異率が一定であれば，分子進化速度は一定となり，いわゆる分子時計が成立するため，これを検定する方式がいくつか提唱されてきた（Felsenstein（1981），Muse and Weir（1992），Tajima（1993））．他方で，数多くあるゲノムの変異の中から，生物の適応進化に本質的に結びついたものを探し出す努力が多くの生物科学者によって積み重ねられてきた．

その一端をトウモロコシに見てみよう．農耕と育種の歴史においては表現型に対する人為的な淘汰圧を通じて，進化的時間から見ると短期間にゲノムが「適応進化」していることが想像さ

[1]東京大学 大学院農学生命科学研究科：〒113–8657 東京都文京区弥生 1–1–1
[2]Bioinformatics Research Center, North Carolina State University, Box 7566, Raleigh, NC 27695–7566, U.S.A.

れる．この意味で，生物進化に関する情報量の多い実験材料とみなすことができる．栽培化されたトウモロコシ（maize）は直立し葉も少なく，多くの実をつけたその容姿は人手により入念に育成されなければ安定して維持することはできない．以前からメキシコに自生する teosinte と自由に交配し，子孫も残すことが知られていた．このため，maize は teosinte から派生して来たのではないかと思う人もいた．確かに雄花である先端部の房状の花序は両者で似ている．が，teosinte は雄花を頂く長い側枝を数多く持つのに対して，maize の側枝は短く，先には雌花である大きな実がついている．teosinte の実は小さく，硬く，食用には全く適さない．

すでに 60 年あまり前の 1938 年，Mangelsdorf と Reevs が teosinte と maize を掛け合わせた交配実験を行い，その子孫の形質の出現頻度を詳細に見ることにより，両者を分けるのは 4 ないし 5 つの遺伝因子であると予想した．ただ，両者の表現型の隔たりは通常観測される進化速度から説明される範囲を大きく越えており，彼らは，トウモロコシは teosinte から派生して来たのではなく，野生種とその近縁種の間で染色体のある領域が大規模に入れ替わったと推測した．これに対して Beadle は，その翌年，4 ないし 5 つの遺伝因子は遺伝子を指していると唱えた．teosinte における数少ない突然変異が奇跡を生み，その昔農民がこの希少な品種を一気に広めたのではないか，と推論した．

近年になって，Doebley et al.（1997）がこの論争にほぼ最終的な決着を下した．彼らは maize にトランスポゾンタギングを施すことにより，長い側枝を数多く持つ変異体 *teosinte branched* 1: *tb*1 を同定した．この突然変異体の形態は teosinte と酷似しており，maize が teosinte からこの遺伝子上の突然変異を含む数少ない突然変異により派生して出来たことを強く支持した．この遺伝子の発現を調べたところ，*tb*1 突然変異体では発現は maize の半分に抑えられていた．

続いて彼らは *tb*1 の遺伝的多様性を見ることにより，農耕の過程でどのように遺伝子に選択圧がかかったか，調べた（Wang et al.（1999））．世界各地から採集した maize および teosinte について *tb*1 遺伝子の各部位について遺伝的多様度を測り，300 塩基の長さの窓を 50 塩基毎にずらしながら，それぞれについて遺伝的多様度を求めた．核を持つ真核生物の遺伝子は，たんぱく質をコードするコード領域の他に，その上流部分には調節領域がある．調節領域は遺伝子の発現を調節するとされている．また，コード領域も，アミノ酸配列に対応するエクソンが mRNA の転写の過程で抜け落ちるイントロンで分断される．コード領域はたんぱく質を生成する重要な機能があるため，もともと teosinte においても多様性は低く，maize が派生して来る中で大きな変化はなかった．機能自体は大きく変わっていないことが示唆される．また 3′ 末端付近の下流部分は本質的でないためか，両者とも多様性が上がる．これに対して，5′ 上流の調節領域では大きな変化が見られる．teosinte ではさまざまな環境に適応して遺伝子の発現の量が多様に調節されているのに対して，maize では一律に過剰発現するよう（人手による）淘汰圧がかかっていることが想像される．

それではこの遺伝子の機能は何か．*tb*1 遺伝子配列と似た配列が既存のデータベースの中にないか，相同性検索を行ったところ，キンギョソウの突然変異体から単離された *cycloidea*（Luo et al.（1996））が釣られてきた．この花は，野生型では一部対称性が崩れて線対称の花弁を持つのに対し，突然変異体の花は円対称である．対称性が崩れた花は蜜を取りやすいことから花粉を運ぶ虫が集まり，結果として集団に固定して行ったのである．野生型では 6 枚の花弁のうち 1 枚が発生の初期の段階で成長を抑えられる．1 枚の花弁の成長が抑えられることにより空き空間に非対称性が生じ，結果としてこのような花になったのである．*cycloidea* 突然変異体では，この花弁の成長を抑える機能が壊され，花弁が円対称に空間を埋めていった．この意味で，teosinte における *tb*1 遺伝子も，側枝の成長を抑えていた．ただし maize では，機能が壊れたのではなく，コード領域の上流にある遺伝子の発現を調節する部分に変異が起き，この機能が倍加されることになったのである．側枝の成長が止められることにより，地下から吸い上げられ

た養分は実に流れて行った．トウモロコシとキンギョソウというまったく異なる植物が似通った配列を共有しており，それらは器官の発生の過程で一部分の成長を抑えるという共通の機能を備えていたのである．

現在のところ，ゲノムレベルでの適応進化の爪痕を検出する際の第一の探索的アプローチとして系統間で分子進化速度を比較する方法が有力視されている．トウモロコシの例から想像されるように，現在では，生物の多様化の多くの部分は調節領域の変異による発現量の変化で説明されるだろう，と信じられている（Doebley and Lukens（1998））．ハワイにおける silversword 群団は 30 種ほどの植物から構成されるが，これらは分子レベルでは極めて近似しているにもかかわらず，形態レベルでは広範に環境適応している．Barrier et al.（2001）はこれらの種におけるシロイヌナズナの花弁の調節遺伝子 APETAL1 と APETALA3，および光合成に関与する構造遺伝子 ASCAB9 と相同な配列を，北アメリカの tarweed におけるそれと比較した．その結果，2 つの調節遺伝子では silversword 群団においてアミノ酸置換を伴う非同義置換の進化速度が大幅に加速されていたことがわかった．これに対して構造遺伝子ではこうした傾向は認められなかった．

RNA ウイルスは世代の長さが 2 日足らずであり，高い突然変異率を持つ．このため，生物進化のプロセスを比較的短期間に観測できる可能性を秘めている．異なる時点から採られたウイルスの配列を比較することにより，進化速度を直接推定することが可能となってくる（Rambaut（2000））．Fitch et al.（1997）はインフルエンザ A 型ウイルスの世界流行の因子の同定と予測を目指して，1984 年から 1996 年にかけて世界各地で採られた，254 の赤血球凝集素遺伝子配列を詳細に調べた．その中で，1992 年以降分子速度が加速されていることが観測された．が，これはワクチン開発のために，この時期から積極的に多様なウイルスを収集し始めたことによる可能性も否定できないとした．Shankarappa et al.（1999）は半年に一度定期的に訪れるエイズ患者から採られたエンベロープ遺伝子を調べたところ，潜伏期間において進化速度はほぼ一定で，速度が鈍ると間もなく発病に至ることを示した．

この他，菌類のうち藻類と共生するもの（地衣体を形成するものやゼニゴケ）と共生関係にないものを比較分析し，共生関係へ移行した後核のリボソーム DNA の進化速度が速くなったことを示した研究（Lutzoni and Pagel（1997））など，進化速度の変動に見られる生物適応の痕跡を見出した研究は枚挙に遑がない．

筆者らはこれまで，分子進化速度が変動する様子を，事前にグループ分けするなど強い制約を置くことなく柔軟に推定する階層モデルを提案した（Thorne et al.（1998），Thorne and Kishino（2002））．そして，分子時計を仮定した解析との対比において，分岐年代推定の頑健性を調べた（Kishino et al.（2001））．さらにこのモデルを複数の遺伝子の解析用に拡張し，遺伝子の速度変動間の相関関係を検出する方法を提案した．シミュレーションを通じて，速度変動の大きさを個々の遺伝子について精度良く推定するためには数多くの配列が必要で，分岐年代の推定については，数多くの遺伝子を解析するときにこの影響がより強く現れることがわかった．ただし，RNA ウイルスのように異なる時点から採られた配列には，分子進化速度に関する情報が多く含まれている．ここではこれらの成果を紹介しつつ，ゲノムデータベースが急速に充実することを念頭におき，階層モデルの今後について言及する．この手法は現在，HIV の起源の頑健な推定や哺乳類の進化と分岐年代の推定など，さまざまなデータに適用され，興味深い結果を生み出している（Cao et al.（2000），Korber et al.（2000））．

2. 分子系統樹の尤度と進化速度の確率変動モデル

マルコフ過程で記述される分子進化の推移速度行列をモデル化することにより，分子系統樹

の尤度が表現される．この尤度には分岐年代と速度行列に関するパラメータが含まれている．分岐年代と速度に事前分布を導入し，これを規定する超パラメータに対して分布を仮定することにより階層モデルが実現される．

2.1 分子系統樹の尤度関数

相同な s 本の配列を比較して系統関係を推定する場合を考える．配列の長さを n とすると，データは

$$\begin{array}{lccccc} \text{種 } 1 & X_{11} & \cdots & X_{1q} & \cdots & X_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{種 } p & X_{p1} & \cdots & X_{pq} & \cdots & X_{pn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{種 } s & X_{s1} & \cdots & X_{sq} & \cdots & X_{sn} \end{array} \tag{2.1}$$

と表現される．ここで X_{pq} は T, C, A or G のどれかで，p 番目の種の第 q 座位の塩基である．$\boldsymbol{X} = (X_{pq})$ を行列表現されたデータ，

$$\boldsymbol{X}_h = (X_{1h}, \ldots, X_{sh})' \tag{2.2}$$

を第 h 座位のデータとする．

進化は座位間で独立とすると，系統樹 T の対数尤度は

$$l(\boldsymbol{\theta}|\boldsymbol{X}) = \sum_{h=1}^{n} \log f(\boldsymbol{X}_h|\boldsymbol{\theta}) \tag{2.3}$$

と表される．ここで $\boldsymbol{\theta}_i$ は進化のプロセスを規定するパラメータである．配列の変化の統計モデルとしては，それが複製の際のミスコピーに起因していることから，マルコフ過程によるモデル化が妥当である．分岐後それぞれの種の配列は独立に進化すると仮定すると，$f(\boldsymbol{X}|\boldsymbol{\theta})$ は

$$f(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{Z_{i_0}} \pi_{Z_{i_0}} \prod_{j \in node(T)\setminus i_0} \sum_{Z_j} P_{Z_{anc(j)},Z_j}(t_{anc(j),j}) \tag{2.4}$$

と簡単に表される．ここで，$node(T)$ は系統樹 T の節を表し，i_0 はその根である．無根系統樹の場合には，任意の節を指定する．$anc(j)$ は j に隣接する祖先となる節である．$P_{xy}(t)$ は時間 t を経た推移確率である．推移速度行列を R と書くと，推移確率行列は $P(t) = \exp(tR)$ として求められる．

塩基置換の統計モデルとしては，すべての塩基に対して等確率で置換することを仮定する Jukes-Cantor モデル（Jukes and Cantor（1969））を基本形として，トランジションとトランスバージョンを区別したモデル（Kimura（1980）），塩基組成の不均一性を考慮に入れたモデル（Felsenstein（1981）），これら 2 つの効果を両方考慮したモデル（Hasegawa et al.(1985), Tamura and Nei（1993））などがあり，また配列内の不均質性を考慮したモデルも開発されている（たとえば Tamura and Nei（1993), Yang（1993））．（2.3）式および（2.4）式はアミノ酸を単位にした場合（たとえば Kishino et al.(1990), Adachi and Hasegawa（1996), Thorne et al.(1996)），あるいは同義置換と非同義置換の対比に注目したコドンの変化をモデル化する場合（たとえば，Miyata and Yasunaga（1980), Muse and Gaut（1994), Goldman and Yang（1994), Nielsen and Yang（1998））にもそのまま適用できる．

2.2 進化速度の事前分布

ここでは，系統樹の形，すなわち分岐の順番は既知であることを仮定する．速さ 1 に規格化された速度行列 R_0 は系統間で異ならず，速さのみが確率変動するモデルを考える．すなわち，

$R = rR_0$（r はスカラー）としたとき，r のみが確率変動するとする．R_0 は経験ベイズを適用し，進化速度に制約を加えないモデルから最尤法により推定する．

進化速度が変動する背景要因として，選択圧の変化を中心とした環境変動，集団の大きさの変動，世代の長さの変動などが考えられる．これらはいずれも自己相関を持って変動することが予想される．そこで，事前分布として速度 $r(t)$ の対数をとったものが簡単な拡散過程に従うとする．

すなわち，$\tilde{r}(t) = \log r(t)$ は正規マルコフ過程で，任意の 2 時点 $t,\ s\ (t > s)$ に対して

$$E\left[\tilde{r}(t)|\tilde{r}(s)\right] = \tilde{r}(s) - \frac{\nu}{2}(t-s)$$
$$V\left[\tilde{r}(t)|\tilde{r}(s)\right] = \nu(t-s)$$

を仮定する．移流項は進化速度の期待値が傾向的に変動させないためにつけたものである．分岐後，速度は 2 系統で独立に変化するとする．

各枝の平均速度は，それをはさむ 2 つの節における速度の平均で近似する．分岐年代を所与とした系統樹の節における速度の対数は，多変量対数正規分布に従う．

分布を規定する超パラメータとして平均速度 μ および拡散係数 ν を持つ．これら 2 つの超パラメータは独立なガンマ分布に従うとする．次項に述べるように時間スケールを規格化し，μ の期待値は 1，標準偏差は 0.5 に設定する．また，ν については強い制約とならないよう，平均，分散ともに 2.0 に設定する．複数の遺伝子を扱う場合には，独立な事前分布を導入する．

2.3 分岐年代の事前分布

分岐年代の事前分布として，種分化と種のサンプリングをモデル化した事前分布も考えられる．これまでのところこれに関して充分信頼できる情報が得られていないことから，ここではこうしたモデル化は避け，できるだけデータの持つ情報を越えた強い制約を課さないよう配慮した．

外群を用意することにより，内群の共通祖先，すなわち根の推定が可能となる．分岐年代の事前分布は，この根に対する事前分布とその他の節に対する事前分布から構成される．根の分岐年代についてはガンマ分布を仮定する．部分的に経験ベイズのアプローチを適用し，その期待値は分子時計を仮定して得られた分岐年代とする．不確実性を考慮に入れるため，期待値の $\frac{1}{2}$ の標準偏差を仮定する．

根以外における内部の節の分岐年代は，緩い事前分布を導入することが重要である．系統樹内の長い経路から順に，内部の節の分岐年代を相対値の形で割り当てて行く．経路が $k-1$ 個の節により k 個の枝に分けられる場合には，それらの長さの相対値 $\pi_i,\ i = 1, \ldots, k$ はディリクレ分布

$$f(\pi_1, \pi_2, \ldots, \pi_k|\alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\pi_i^{\alpha_i - 1}$$

に従うとする．ここで，$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1$ とする．これは $k-1$ 個の一様乱数により内部の節の相対的な位置を決めることと同等である．

図 1 に見られる例では，まず経路 $0 \to 6 \to 9 \to 10$ 上の 2 つの節 6 および 9 の位置を決める．続いて，経路 $3 \to 7 \to 9$ 上の節 7 を決め，同様に，経路 $5 \to 8 \to 10$ 上の節 8 を決める．こうして得られる事前分布は経路の取り方に依らずに決まることが容易にわかる．

さらに，化石などから，ある節における分岐年代の下限，上限，あるいは両方の証拠がわかっている場合がある．この場合は，上の事前分布をこの区間に制限することにより，この情報を加味した事前分布が得られる．

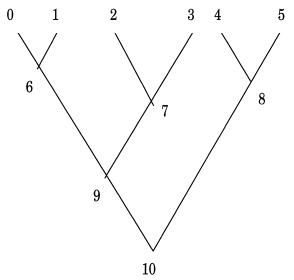


図 1. 根つき系統樹と分岐年代の事前分布（本文参照）.

2.4 二段階方式による近似モデルとメトロポリス−ヘイスティングアルゴリズム

尤度関数の形から，塩基組成と各枝における各種塩基置換数が，分子進化のさまざまなモデルにおいて十分統計量であることがわかる．後述するように，我々の提案する方法は二段階からなっており，第一段で枝毎の塩基置換数を最尤推定し，その分散共分散行列を Fisher 情報行列から評価する．第二段ではこの推定量が多変量正規分布に従うとみなし，メトロポリス−ヘイスティング法（Metropolis et al.(1953), Hastings（1970)）で分岐年代および各枝の速度，速度変化の大きさを規定する超パラメータの事後分布を求める．第一段とは切り離されているところに特徴がある（Thorne et al.(1998)）．塩基置換，アミノ酸置換およびコドン変化のモデルは第一段にのみ影響し，第二段の取り扱いは同一である．

すなわち，内部の節における速度を $\boldsymbol{r}$，分岐年代を $\boldsymbol{t}$，各枝における期待塩基置換数を $\boldsymbol{b}$ とおくと，

$$p(\boldsymbol{t},\boldsymbol{r},\nu|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{b})p(\boldsymbol{r}|\boldsymbol{t},\nu)p(\boldsymbol{t})p(\nu)}{p(\boldsymbol{X})} \tag{2.5}$$

となる．メトロポリス−ヘイスティング法の各ステップの詳細はここでは触れず，モデルの紹介をするに止めるが，$p(\boldsymbol{X}|\boldsymbol{b})$ の比の扱いでは，

$$p(\boldsymbol{X}|\boldsymbol{b}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{b}-\hat{\boldsymbol{b}})' H^{-1}(\boldsymbol{b}-\hat{\boldsymbol{b}})\right] \tag{2.6}$$

と Laplace 近似する．$\hat{\boldsymbol{b}}$ は各枝の長さを自由パラメータとしたときの最尤推定値で，H は対数尤度（(2.3）式）の Fisher 情報行列により求められる．

2.5 遺伝子間の共進化の検出

複数の遺伝子は世代の長さを共有する．またこれらの遺伝子が互いに関連した機能を持っている場合には，外界から同方向の選択圧を受けて来ていることが推察される．その結果，これらの遺伝子は同様の系統で速度が加速される，あるいは逆に同時に減速される，という共進化を経験することになる．

我々が開発した複数遺伝子の階層モデルでは，現在のところ速度変化の事前分布として遺伝子間で独立な確率過程を採用している．相関を許した多変量の確率過程を導入することにより，超パラメータの事後分布の形で遺伝子間の共進化を検出することができよう．ただ，数多くの遺伝子を同時解析するときには，その間の相関を記述する超パラメータの数は遺伝子数の 2 乗のオーダーで膨らんで行く．ゲノムデータベース解析を睨んだ実際的な配慮から，超パラメータの事後分布として共進化を検出する方式を取らず，遺伝子間で独立を仮定した事前分布に対する速度の事後平均を遺伝子間で比較し，相関の有意性を見る（Thorne and Kishino（2002)）.

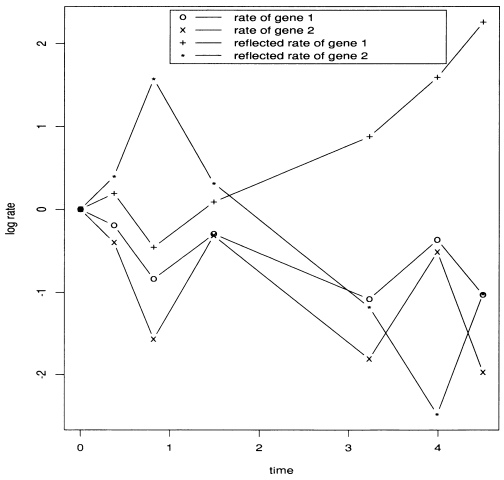


図 2. 反射法による無相関速度変化のシミュレーション.

2 遺伝子における各節での速度の事後平均 $\overline{r}_{mj}$ $(m = 1, 2,\ j \in node(T))$ を求め，順位相関を取る．無相関の帰無仮説の下での分布は，速度変化の確率変動と推定誤差による不確実性を踏まえていなければならない．ここでは，速度変化の方向を無作為化することにより分布を得る．すなわち，帰無仮説の下での各節における速度の事後平均 $\overline{r}'_{mj}$ を

$$\overline{r}'_{m,i_0} = \overline{r}_{m,i_0}$$
$$\overline{r}'_{m,j} = \overline{r}'_{m,anc(j)} + W_{mj}(\overline{r}_{m,j} - \overline{r}_{m,anc(j)}) \quad (j \in node(T) \setminus i_0)$$

により生成する．W_{mj} は $1,\ -1$ をそれぞれ確率 $\frac{1}{2}$ でとる互いに独立な確率変数である．

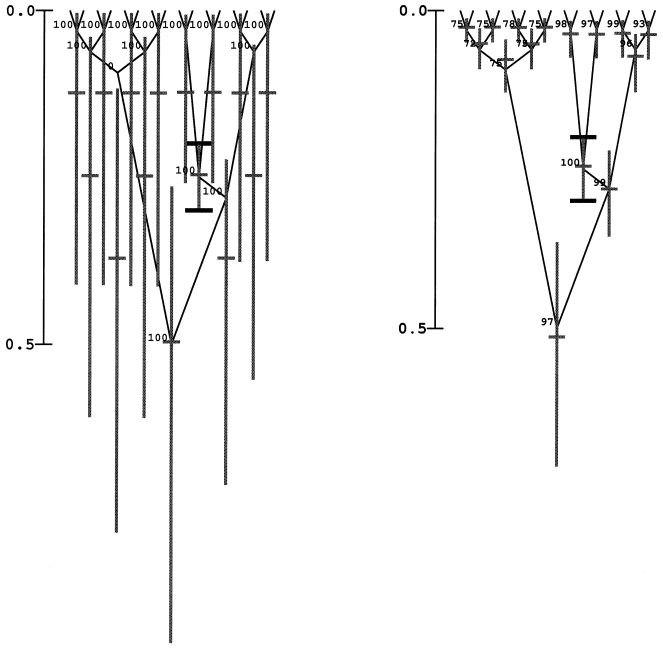
図 2 では，ある経路に沿った内部の節における 2 つの遺伝子の事後平均速度を模式的に示しており，正の相関を持って確率変動していることが示唆される（“○”印と “×”印）．これに対して，“+”印と “∗”印は帰無仮説の下での順位相関の分布を出すためのシミュレーションの 1 実現値で，変動の幅を保持しながら変化の向きをランダムに与えることにより 2 遺伝子間の相関を断ち切っている様子が見て取れる．

3. シミュレーションによる性能評価

ここで紹介した進化速度の確率変動を記述する階層モデルの性能をシミュレーションを通じて評価した．乱数により進化速度を確率変動させ，Jukes-Cantor モデルにより長さ 1000 の配列を生成した．その結果，分子時計が成り立つときにおいても，進化速度の確率変動を記述する超パラメータを導入することによる精度の低下は小さく，速度が変化するときは，階層モデルは分子時計を仮定するモデルに比し，はるかに頑健に分岐年代を推定することが示された（Kishino et al.（2001））．ここでは，分岐年代の事前分布に対する結果の感受性と，複数の遺伝子を解析するときの情報の付加について報告する（Thorne and Kishino（2002））．

3.1 分岐年代の事前分布と事後分布

確率変動の自己相関を表現する ν の期待値，標準偏差をそれぞれ 2.0 に設定し，100 回のシミュレーションを行った．分岐年代の事前分布に対する感受性テストであるため，ν の事前分



a) 事前分布　　b) 事後分布

図 3. 最近に分岐が集中した系統樹における分岐年代の事前分布と事後分布（Kishino et al.(2001)）.

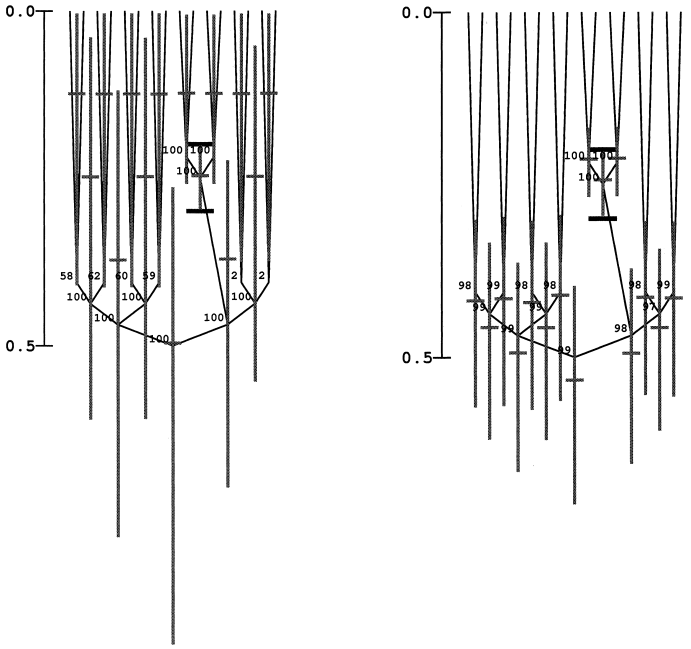
布も期待値，標準偏差がそれぞれ 2.0 であるとして解析を行った．分岐年代の事前分布はディレクレ分布により生成されるが，化石情報などの制約条件が加わったときは，その形を目に見える形で表現するのは難しい．(2.6) 式を定数で置き換えると，データが情報を持たない場合の事後分布として事前分布を求めることができる．図 3 (a) は，最近に分岐が集中した系統樹と分岐年代に対する事前分布である．16 の配列と 1 つの外群からなる．

横棒は事後平均の平均を表し，縦棒は同じく 95%信頼区間の上限値および下限値の平均を示している．節の横に添えられた数字は，この 95%信頼区間が真値をカバーする確率をパーセントで示したものである．内部の節の 1 つに太い横棒で挟まれた制約条件がかかっており，ここにおいては分岐年代の事前分布はこの区間に収まっている．その他の節はこの部分の影響をほとんど受けず，大きな分散を持ちながら平均的には節が各経路を等分している様子が読み取れる．これに対して図 3 (b) は事後分布を示している．事前分布の影響から解き放たれて，真値をほぼ偏りなく推定していることがわかる．

逆に，分岐年代が過去の 1 時点近辺に集中した場合のシミュレーションが図 4 である．同じく節の分岐年代の事前平均は各経路を等分し，事後分布は真値をほぼ偏りなく復元している．

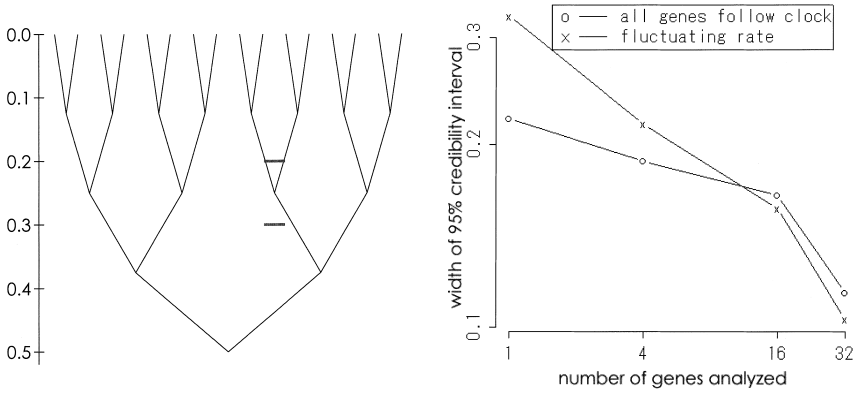
3.2 複数遺伝子モデルの評価

2 つの場合について，複数の遺伝子を解析することによりどの程度情報が付加されるか，調べる（Thorne and Kishino（2002））．まずは化石情報を想定し，ある節において分岐年代の上限と下限が与えられている場合（図 5 (a)）である．解析する遺伝子数が 1, 4, 16, 32 のそれぞれについて，進化速度が一定の場合と確率変動する場合のそれぞれについてシミュレーション



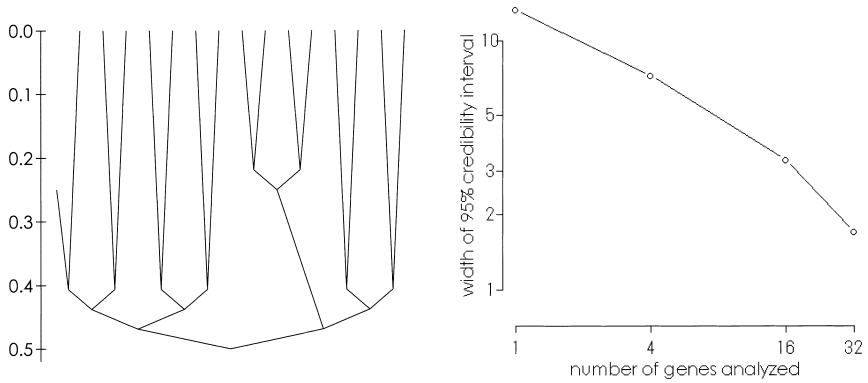
a) 事前分布　　b) 事後分布

図 4. 過去の 1 時点に分岐が集中した星型系統樹に近くなった場合の分岐年代の事前分布と事後分布（Kishino et al.（2001））.



(a) 系統樹　　(b) 解析する遺伝子数と根の95%信頼区間の幅

図 5. 内部の節に幅を持った制約条件がかかった系統樹を推定する場合の解析遺伝子数と分岐年代の推定精度（Thorne and Kishino（2002））.



(a) 系統樹　　(b) 解析する遺伝子数と根の95%信頼区間の幅

図 6. 異なる時点からの配列を解析する場合の解析遺伝子数と分岐年代の推定精度（Thorne and Kishino (2002)）.

を行った．横軸に遺伝子数，縦軸に根の分岐年代の 95%事後信頼区間の幅をとり，両対数でプロットを行った．図 5 (b) は，ほぼ直線が当てはまることから，遺伝子数の負のべき乗に従って事後標準偏差が小さくなって行くことが示唆される．ただ，参照点となる節において幅を持った事前情報のみが得られることから，そのスピードは遺伝子数の平方根には比例しない．

第二はウイルス進化や古代 DNA の解析を想定し，ある配列が異なる時点により得られたことを仮定する場合（図 6 (a)）である．ここでは進化速度が一定の場合について，遺伝子数を増やす効果を調べた（図 6 (b)）．現時点以外のある節の年代が曖昧さなく既知となっているため，根の 95%信頼区間の幅は遺伝子数の平方根に逆比例している様子が見て取れる．

4. 考察：ゲノムデータベース解析と階層モデル

ゲノムの多様化する機構の推測において，最も興味深いテーマのひとつに遺伝子や機能の得失，およびその背景要因の推定がある．比較ゲノムにより遺伝子の得失の履歴が推定できるようになってきた（Pellegrini et al. (1999), Mizuno et al. (2001)）．原核生物においては水平伝播によりしばしば遺伝子を獲得する可能性があるが，真核生物においては遺伝子重複を通じた遺伝子の獲得の主たるものと信じられている（Ohno (1970)）．そうした証拠は，しばしばゲノム内に似た配列が複数存在することから推察することができる．

遺伝子重複により同一の遺伝子が 2 つできると，このままの状態では過剰発現し，しばしばそれは有害に働くであろう．こうしたことから一方の遺伝子は機能的な制約を解かれ，自由に変異し，やがては遺伝子として機能しなくなり，偽遺伝子となって行く．重複遺伝子がそのまま保持されるための要因としては，変異の後たまたま新たな機能が加わる，あるいは複数のコピーを持つことに対して正の淘汰圧がかかる，といったメカニズムが考えられていた（Ohta (1987, 1988); Nowak et al.(1997)）．新たな機能の獲得（neo-functioning）に対して，Force et al. (1999) および Lynch and Force (2000) は，部分機能化（subfunctionalization）による重複遺伝子の維持を提唱し，集団遺伝学的な理論解析を行った．調節領域を複数持つ遺伝子が重複した場合，コード領域の変異は多くの場合死滅するが，調節領域の変異は時間特異的，あるいは組織特異的な発現へと導かれる．これによると，重複後の両遺伝子の寿命は長くなる．

Zhang et al.（1998）は，EDN 遺伝子から派生した ECP 遺伝子が，遺伝子重複後新たな機能を獲得する過程で比較的短期間に，正の淘汰圧を受けている証拠を示した．エオシン好性陽性たんぱく（eosinophil cationic protein: ECP）とエオシン好性神経性毒素（eosinophil-derived neurotoxin: EDN）はともに好エオシン白血球固有の大型顆粒中に存在する．EDN は RNA 分解酵素に対して強い触媒作用を持ち，神経毒性を持つ．これに対して ECP は，触媒作用は弱いが病原体に対する強い毒性を持つ．分子系統学的な解析により，新世界ザルと旧世界ザルが分かれ，旧世界ザルからヒト科が分岐する以前に，EDN 遺伝子の祖先遺伝子から ECP 遺伝子が遺伝子重複により誕生したことが判明した．そして，非同義置換と同義置換の数を系統毎に推定することにより，遺伝子重複後ヒト科が分岐する以前に限ってアミノ酸置換を伴う非同義置換がアミノ酸置換を伴わない同義置換に比し有意に多く観察されることを突き止めた．さらにアミノ酸の変異をさらに詳細に調べたところ，アルギニンがこの間に著しく増加しており，この変化が病原体に対する強い毒性の機能獲得に結びついたと推測した．遺伝子重複の中でも，植物においてしばしば観測されるゲノムの倍数化は特別の意味を持つ．倍数化した系統や種は環境に対する適応力が大きく，分布範囲も広いことが多い．Cronn et al.（1999）はワタの異質倍数体における 16 の部位を解析し，重複後これらが互いに独立に進化していること，しかも通常の遺伝子重複と異なり，重複遺伝子の一方において進化速度の加速が観察されないことを見出した．ゲノムの倍数化には小領域の重複を越えた安定化作用が働く要因があるのかも知れない．

Lynch and Conery（2000）はこうした事例研究を超えて，ゲノムレベルで解析を行い，重複遺伝子の生成と死滅について推測を行った．彼らは，ゲノムの整備された数種の真核生物について，機能を持っていると思われるアミノ酸配列に相同性検索（Altschul et al.（1997））をかけることにより数多くの重複遺伝子対をはじき出した．そして，比較的中立的であると期待される同義置換を時計に見立て，この遺伝子対のデータに重複後の非同義置換の速度変化を記述した簡単な微分方程式を当てはめた．その結果，どの生物においても，機能を持っているときは平均的に非同義置換の速度は同義置換の速度の 5%程度に抑えられており，重複後一方の遺伝子が機能的制約を解かれて非同義置換の速度が同義置換の 8 割程度まで上昇することが示唆された．さらに，同義置換で計った重複遺伝子対の年齢分布から，遺伝子重複により機能的制約を解かれた遺伝子が消失するまでの半減期を 300 万年から 700 万年程度と推定した．さらに近接した重複遺伝子対の数から，100 万年のうちに 100 個から 500 個中 1 個の遺伝子が重複する，と推定した．間接的な観測に基づく大雑把な解析であるが，全ゲノムを用いてその進化の様相をマクロ的に捉えた点で注目される．

本稿では分子進化の階層モデルとその性能を概観した．これまでのところ整備されているのは比較的少数の遺伝子を解析する分析枠組である．種々の生物の全ゲノム配列が次々に解読されるにつれ，数多くの相同な配列を自在に解析し，比較ゲノム解析を行うことが可能になってきた．そこにおいては，超パラメータを通じた事前分布は，ゲノムにおける遺伝子の特性の分布という実質的な意味を持つ．従って階層モデルは，ゲノムデータベースに格納された遺伝子の間のある特性の分布を記述するのに優れたモデルといえる．事前分布のモデリングが主な役割を担って行くであろう．

我々の二段階方式に対して，こうした近似をせずに，完全な形で事後分布を推定する方法もその後提案されている（Huelsenbeck et al.（2000））．まだ原因は不明であるが，マルコフ連鎖モンテカルロ法の収束性は芳しくないようである．ところで，進化速度に制約を課すことなく得られた系統樹は，それ自体意味を持つことに注意したい．これを通して，進化速度が変容する様をさまざまな角度から検討しながら，妥当な事前分布をモデル化することが可能となってくる．二段階方式はこのトータルな作業を自然に具体化したもので，特に複雑な集団構造を持

つ対象の集団遺伝学的構造を推定するとき，あるいは複雑かつ巨大なゲノムデータベースをモデル化するときに有効なアプローチとなるであろう．

謝　辞

本研究の一部は文部科学省科学研究費補助金，日本学術振興会日米科学協力事業，米国科学財団研究費交付金の助成を受けた．

参 考 文 献

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA, *Journal of Molecular Evolution*, **42**, 459–468.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Research*, **25**, 3389–3402.

Barrier, M., Robichaux, R. H. and Purugganan, M. D. (2001). Accelerated regulatory gene evolution in an adaptive radiation, *Proc. Nat. Acad. Sci. U.S.A.*, **98**, 10208–10213.

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

Cronn, R. C., Small, R. L. and Wendel, J. F. (1999). Duplicated genes evolve independently after polyploid formation in cotton, *Proc. Nat. Acad. Sci. U.S.A.*, **96**, 14406–14411.

Doebley, J. and Lukens, L. (1998). Transcriptional regulators and the evolution of plant form, *Plant Cell*, **10**, 1075–1082.

Doebley, J., Stec, A. and Hubbard, L. (1997). The evolution of apical dominance in maize, *Nature*, **386**, 485–488.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, **17**, 368–376.

Fitch, W. M., Bush, R. M., Bender, C. A. and Cox, N. J. (1997). Long term trends in the evolution of H(3) HA1 human influenza type A, *Proc. Nat. Acad. Sci. U.S.A.*, **94**, 7712–7718.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y-L. and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations, *Genetics*, **151**, 1531–1545.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Molecular Biology and Evolution*, **11**, 725–736.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, **22**, 160–l74.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97–109.

Huelsenbeck, J. P., Larget, B. and Swofford, D. L. (2000). A compound Poisson process for relaxing the molecular clock, *Genetics*, **154**, 1879–1892.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules, *Mammalian Protein Metabolism* (ed. H. N. Munro), 21–32, Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences, *Journal of Molecular Evolution*, **16**, 111–120.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, New York.

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.

Kishino, H., Thorne, J. L. and Bruno, W. J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution, *Molecular Biology and Evolution*, **18**, 352–361.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S. and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains, *Science*, **288**, 1789–1796.

Luo, D., Carpenter, R., Vincent, C., Copsey, L. and Coen, E. (1996). Origin of floral asymmetry in antirrhinum, *Nature*, **383**, 794–799.

Lutzoni, F. and Pagel, M. (1997). Accelerated evolution as a consequence of transition to mutualism, *Proc. Nat. Acad. Sci. U.S.A.*, **94**, 11422–11427.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes, *Science*, **290**, 1151–1155.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization, *Genetics*, **154**, 459–473.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1092.

Miyata, T. and Yasunaga, T. (1980). Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application, *Journal of Molecular Evolution*, **16**, 23–36.

Mizuno, H., Tanaka, Y., Nakai, K. and Sarai, A. (2001). ORI-GENE: Gene classification based on the evolutionary tree, *Bioinformatics*, **17**, 167–173.

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome, *Molecular Biology and Evolution*, **11**, 715–724.

Muse, S. V. and Weir, B. S.(1992). Testing for equality of evolutionary rates, *Genetics*, **132**, 269–276.

Nielsen, R. and Yang Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, *Genetics*, **148**, 929–936.

Nowak, M. A., Boerlijst, M. C., Cooke, J. and Smith, J. M. (1997). Evolution of genetic redundancy, *Nature*, **388**, 167–170.

Ohno, S. (1970). *Evolution by Gene Duplication*, Springer, Berlin.

Ohta, T. (1987). Simulating evolution by gene duplication, *Genetics*, **115**, 207–213.

Ohta, T. (1988). Time for acquiring a new gene by duplication, *Proc. Nat. Acad. Sci. U.S.A.*, **85**, 3509–3512.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Nat. Acad. Sci. U.S.A.*, **96**, 4285–4288.

Rambaut, A.(2000). Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies, *Bioinformatics*, **16**, 395–399.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L. and Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection, *Journal of Virology*, **73**, 10489–10502.

Tajima, F.(1993). Simple methods for testing the molecular evolutionary clock hypothesis, *Genetics*, **135**, 599–607.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, *Molecular Biology and Evolution*,

10, 512–526.

Thorne, J. L. and Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data, *Systematic Biology* (in press).

Thorne, J. L., Goldman, N. and Jones, D. T. (1996). Combining protein evolution and secondary structure, *Molecular Biology and Evolution*, **13**, 666–673.

Thorne, J. L., Kishino, H. and Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution, *Molecular Biology and Evolution*, **15**, 1647–1657.

Wang, R. L., Stec, A., Hey, J., Lukens, L. and Doebley, J. (1999). The limits of selection during maize domestication, *Nature*, **398**, 236–239.

Yang, Z.(1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, *Molecular Biology and Evolution*, **10**, 1396–1401.

Zhang, J., Rosenberg, H. F. and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes, *Proc. Nat. Acad. Sci. U.S.A.*, **95**, 3708–3713.

Bayesian Hierarchical Model of Rate of Molecular Evolution

Hirohisa Kishino

(Graduate School of Agriculture and Life Sciences, University of Tokyo)

Jeffrey L. Thorne

(Bioinformatics Research Center, North Carolina State University)

In the evolution process, biological organisms leave traces of diversification and adaptation to a genome in the form of evolutionary rate and its change. This paper first examines how the inference of the rate change plays an important role in evolution research, with a few examples from recent works on *tb*1 of domesticated maize and the selection pressure on it, diversified species in the Hawaiian silversword alliance and accelerated rate in regulatory genes, the viral adaptation process to the hosts, and the fate of gene duplication. It then introduces our hierarchical model that describes stochastic change of evolution rate, and briefly evaluates the performance by simulation. Finally, it discusses the possibility of hierarchical models in genome database analysis.

Key words: Stochastic change of evolutionary rate, hierarchical model, Markov chain Monte Carlo (MCMC), detection of correlated evolution, models for multiple genes, genome database analysis.