The Japan Statistical Society

日本統計学会誌
第34巻，第2号，2005年3月
209頁～212頁

分子系統樹推定における統計科学の役割

長谷川　政　美

Role of Statistical Science in Molecular Phylogenetic Inference

Masami Hasegawa

生物学の多くの問題において，生物の系統関係を知ることは議論の出発点である．ゲノムデータの存在が当たり前になったポストゲノムの時代において，分子系統樹推定のための統計的方法はますます重要になってきている．ここでは，分子系統樹推定においてDNA塩基置換がどのようにモデル化されているかについて紹介する．

Phylogenetic relationships among organisms are essentially relevant to most of the biological problems. In the post-genome era, statistical methods for molecular phylogenetic inference are becoming important. I will review how the process of nucleotide substitutions is modeled in the molecular phylogenetic inference.

はじめに

DNAの塩基配列やそれにコードされている蛋白質のアミノ酸配列のデータから，生物進化の系統樹を推定する分子系統樹法は，生物科学のあらゆる分野で重要なツールになっている．現在生きている生物は長い進化の歴史の産物であるから，生物学のあらゆる問題において進化的な視点が重要である．進化的な議論の出発点は，生物の系統関係がどのようになっているかということであり，分子系統樹推定はそれを明らかにする上で重要である．たとえ生物進化を意識しない研究であっても，分子系統樹法が重要な役割を果たすこともある．例えばHIVウイルスや西ナイル熱ウイルスの伝播経路解明のような疫学的な研究にも，分子系統樹法が決定的に重要な役割を果たしているのである（Anderson et al., 1999; Metzker et al., 2002）．

最尤法による分子系統樹推定と置換過程のモデリング

進化におけるDNA塩基や蛋白質アミノ酸の置換は，確率過程とみなすことができる．分子レベルで変異が起こるためには，まず生殖細胞のなかのDNA上で突然変異が生じることが必要であり，これはその名前が示すように確率的な現象である．しかしながら，突然変異はあくまでも個体レベルの現象であり，進化は1つの生物種全体が変化していくことである．個体レベルで起こった突然変異が進化に寄与するためには，突然変異遺伝子が子孫に受け継がれ，そのような遺伝子をもった子孫が増えることによって，集団全体に広がることが必要である．このことを突然変異遺伝子の集団への固定というが，この過程でも偶然的な要素が重要であることが明らかになってきた（木村, 1986）．従って，そのような進化の結果として生成された現生生物の分子配列データから進化の歴史を再構築することは，統計的推測の問題になる．

現在広く用いられるようになってきた最尤法による分子系統樹推定法は，Felsenstein

NII-Electronic Library Service

The Japan Statistical Society

(1981) によってはじめて定式化された．これはDNA塩基配列データを解析するためのものであったが，ここで用いられた塩基置換モデルは塩基組成の偏りだけを考慮した簡単なものであり，現在Felsenstein81モデルと呼ばれている．その後，塩基組成の偏りのほかに，A⟷G, T⟷C 間の transition と A,G⟷T,C 間の transversion の rate の違いを考慮した Hasegawa, Kishino and Yano (1985) モデル (HKY モデル)，可逆性だけを取り入れたもっと一般的な遷移行列を用いた General Time-Reversible (GTR) model (Rodriguez et al., 1990) などが多くの解析ソフトにインプリメントされている (Yang, 1997)．また，座位ごとの rate の違いを discrete Γ 分布 (Yang, 1996) で近似することも一般的である．さらに，蛋白質における特定のアミノ酸座位の進化的な変わりやすさは，分子内のまわりの状況に依存するはずであり，そのようなことをモデル化した上で，分子系統樹の推定を行うこともできる (Penny and Hasegawa, 2001)．

1980年代に Felsenstein が最初に最尤法による分子系統樹推定法を開発した当時は，この方法では計算時間が膨大にかかるため，一般には実用的な方法とは見なされなかった．Hasegawa et al. (1985) は，この方法をはじめて生物学の実際問題に適用したが，分子系統学の分野では1980年代を通じて最尤法の重要性はあまり評価されなかった．そのため，現在最尤法による系統樹の信頼集合を求める方法として広く使われている Kishino-Hasegawa (1989) 検定の論文も，最初にこの論文を投稿したこの分野の代表的な Journal からは，最尤法は実用的な方法でないからこの方法も生物学の実際問題を解くには役に立たないだろう，という理由でリジェクトされてしまった．

1990年代に入ると，コンピュータの能力の飛躍的な進歩と実用的なソフトウエアの開発もあって，最尤法による解析法も分子系統学の分野で次第に認められるようになってきた (長谷川・岸野, 1996; 岸野・浅井, 2003; Felsenstein, 2003; Nielsen, 2004)．これまで，塩基配列のデータしか扱えなかったが，アミノ酸配列データから最尤法によって分子系統樹を推定する方法が Kishino, Miyata and Hasegawa (1990) により開発され，MOLPHY プログラムパッケージにインプリメントされた (Adachi and Hasegawa, 1996a)．アミノ酸の遷移行列は多くの実データから推定したものを用いるのが現実的である．核コードの蛋白質，ミトコンドリア・ゲノムにコードされた蛋白質，葉緑体ゲノムにコードされた蛋白質はそれぞれに特徴があり，それぞれに対応したモデルが開発されている (Jones, Taylor and Thornton, 1992; Adachi and Hasegawa, 1996b; Adachi et al., 2000)．

アミノ酸レベルではこのような実データに基づいたモデルが有用であるが，現在アミノ酸配列データのほとんどは，DNA の塩基配列データを遺伝コード表に基づいてアミノ酸配列に翻訳して得られたものである．コード表では3連塩基 (コドン) が1つのアミノ酸に対応するが，コドンが縮退しているため DNA の塩基置換のなかにはアミノ酸を変えないものもある．アミノ酸レベルの解析では，このような同義置換の情報が取り入れられないために，情報のロスがある．このために，特に近縁な生物種間の系統樹推定では塩基レベルの解析が望ましい．ところが，一般に行なわれている塩基レベルの解析には問題がある．

現在広く使われているプログラムに MODELTEST (Posada and Crandall, 1998) がある．これにはさまざまな塩基置換モデルがインプリメントされていて，ユーザーは AIC を使ってその中から自分の扱っているデータに最も適合したモデルを選択し，そのモデルを用いて別のプログラム PAUP* (Swofford, 1996) で分子系統樹解析をすることができるようになっている．MODELTEST は儀式的に使われているが，問題はここにインプリメントされているモデルがいずれも，蛋白質遺伝子の進化を近似するには現実から離れすぎているということである．非現実的なモデルのセットの中から，ベストのものを選び出してもあまり意味はないのである．

NII-Electronic Library Service

The Japan Statistical Society

蛋白質をコードしている遺伝子は，3連塩基コドンを単位として構成されており，コドン内のそれぞれの塩基の置換は決して独立ではない．ところが，MODELTESTにインプリメントされているモデルは，いずれも独立性を仮定している．アミノ酸に対応したコドンは61種あるので，本来は61×61の遷移行列を扱うコドン置換モデルを用いることが望ましい．Yang, Nielsen and Hasegawa (1998) は，transitionとtransversionの間のrateの違い，同義置換と非同義置換（アミノ酸の変化を伴うコドン置換）の間のrateの違い，アミノ酸間の物理化学的な距離による非同義置換の間のrateの違いなどを考慮に入れたコドン置換モデルを開発した．これはPAMLにインプリメントされているが，このようなモデルを採用することにより，実データとのあてはまりが格段に向上することが示される（Cao and Hasegawa, in preparation; Sasaki et al., in press）．

モデルはあくまでも現実の過程を近似的に表現するものに過ぎないので，限られたデータを非常にうまく近似したモデルであっても，データ量が増えてくると現実とのずれが次第に目立つようになってくる．従って，常に最新の知見を取り入れてモデルをより現実に即したものにする努力を続けていくことが必要である．

分子系統樹法のプログラムのユーザーのなかには，解析法の内容をよく理解しないで使っているひとも多い．上で述べたMODELTESTとPAUP*を組み合わせた使い方では，多くのモデルをAICの規準で比較した上で，データに最も適合したモデルを用いて系統樹推定ができるということで，ユーザーの多くはこれでよいのだという自己満足に陥る傾向がある．問題はここで用意されているモデルがいずれも著しく現実から離れたものであるということである．実質科学で用いられるデータ解析法の開発に携わる統計科学の研究者は，一般のユーザーがこのような誤った自己満足に陥らないように，自分の方法の長所とともにその限界もまた明らかにしておくことも必要であろう．

参 考 文 献

[1] Adachi, J. and M. Hasegawa (1996a). MOLPHY: Programs for molecular phylogenetics ver. 2.3, Computer Science Monographs, No. 28. Institute of Statistical Mathematics, Tokyo.

[2] Adachi, J., and M. Hasegawa (1996b). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42**, 459-468.

[3] Adachi, J., P. Waddell, W. Martin and M. Hasegawa (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.* **50**, 348-358.

[4] Anderson, J. F., T. G. Andreadis, C. R. Vossbrinck, S. Tirrell, E. M. Wakem, R. A. French, A. E. Garmendia, H. J. Van Kruiningen (1999). Isolation of West Nile Virus from Mosquitoes, Crows, and a Cooper's Hawk in Connecticut. *Science* **286**, 2331-2333.

[5] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376.

[6] Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

[7] 長谷川政美，岸野洋久（1996）．分子系統学，岩波書店．

[8] Hasegawa, M., H. Kishino, and T. Yano (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160-174.

[9] Jones, D. T. and W. R. Taylor, and J. M. Thornton (1992). The rapid genaration of mutation data matrices from protein sequences. *CABIOS* **8**, 275-282.

[10] 木村資生（1986）．分子進化の中立説，紀伊國屋書店．

[11] 岸野洋久，浅井　潔（2004）．生物配列の統計：核酸・タンパクから情報を読む，岩波書店．

[12] Kishino, H. and M. Hasegawa (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170-179.

[13] Kishino, H., T. Miyata, and M. Hasegawa. (1990). Maximum likelihood inference of protein phylogeny, and

NII-Electronic Library Service

The Japan Statistical Society

the origin of chloroplasts. *J. Mol. Evol.* **31**, 151-160.

[14] Metzker, M. L., D. P. Mindell, X.-. Liu, R. G. Ptak, R. A. Gibbs, and D. M. Hillis (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl. Acad. Sci. USA* **99**, 14292-14297.

[15] Nielsen, R. (2004). *Statistical Methods in Molecular Evolution*, Springer.

[16] Penny, D. and M. Hasegawa (2001). Covarion model of molecular evolution. *Encyclopedia of Genetics*, edited by S. Brenner and J.H. Miller (Academic Press, San Diego) 473-477.

[17] Posada, D. and K. A. Crandall (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818 .

[18] Rodriguez, F. J., J. L. Oliver, A. Marin, and J. R. Medina (1990). The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**, 485-501.

[19] Swofford, D. (1996). PAUP*, ver. 4, Sinauer Associates, Sunderland, Mass.

[20] Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *TREE* **11**, 367-372.

[21] Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**, 555-556.

[22] Yang, Z., R. Nielsen, and M. Hasegawa (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600-1611.

NII-Electronic Library Service