様式 C -19

科学研究費補助金研究成果報告書

平成 21 年 5 月 29 日現在

研究種目：若手研究（B）
研究期間：2006 -2008
課題番号：18770216
研究課題名（和文） 複数の遺伝子を用いる系統樹推定の統計的手法開発
研究課題名（英文） Development of statistical methods for phylogeny reconstruction using multilocus sequence data

研究代表者
氏 名（アルファベット）：徐 泰健（Seo, Tae -Kun）
所属機関・所属部局名・職名 ： 東京大学・大学院農学生命科学研究科・特任助教
研究者番号 ： 60401189

研究成果の概要：複数の遺伝子の塩基配列を結合せずに、個別に解析する新しい分子系統樹推定法を開発した。また遺伝子と配列サイトを再抽出する 2 段階サンプリング方式を提案し、この方法の妥当性を哺乳類の核遺伝子の解析で検証した。新しい方法の適用により、遺伝子ごとに違う進化的特徴を保ちながら系統樹の不確実性を測ることができるようになった。分子進化モデルと分岐年代推定法の開発も並行して行い、哺乳類の分岐年代とミトコンドリアタンパク質コード遺伝子の同義置換・非同義置換の変動パタンを推定した。

交付額

（金額単位：円）

	直接経費	間接経費	合 計
18 年度	1,400,000	0	1,400,000
19 年度	1,200,000	0	1,200,000
20 年度	900,000	270,000	1,170,000
年度			
年度			
総 計	3,500,000	270,000	3,770,000

研究分野：生物科学
科研費の分科・細目：進化生物学
キーワード：分子系統樹、距離行列法、分子進化

１．研究開始当初の背景

近年、ゲノム情報量の増加と塩基配列決定技術の進歩により、複数の遺伝子を用いた分子進化研究が可能になっている。 その一方、適切な解析手法の開発は遅れており、複数の遺伝子を単純に繋ぎ合わせる「Sequence Concatenation」アプローチが主に使われていた。 繋ぎ合わせた塩基配列を単一遺伝子とみなすこの方法に対して、数多い問題点が指摘されていた。例えば、遺伝子組換え・水平伝播などにより遺伝子ごとに進化の履歴が異なる場合がある。このバリエーションにつ

いて十分考慮せずに塩基配列を単純に結合し、間違った系統樹を得たケースが報告されていた。

配列進化モデルに基づいた分子系統樹の推定法は、大きく分けて最尤法・ベイズ法・距離行列法の 3 種類がある。このうち最尤法・ベイズ法に対して、いくつかの複数遺伝子解析方法が提案されているものの、距離行列法に対しては適切な解析手法がまだ知られていなかった。

２．研究の目的

本研究の主な目的は「Sequence Concatenation」に依存しない、系統樹推定の距離行列法の開発にあった。最尤法とベイズ法に比べて距離行列法には、計算時間が短いというメリットがある。適切なアルゴリズムの開発により、より迅速な系統樹解析ができることを目指した。

３．研究の方法

繋ぎ合せた塩基配列から１つの距離行列を求める代わりに複数の遺伝子から各々の距離行列を求め、これらを処理する方法を以下のように考えた。 i 番目の遺伝子から得られた距離行列を$\boldsymbol{\delta}_i$と表し、$\boldsymbol{\delta}_i$をランダム変数と仮定する。その平均（$\boldsymbol{\Delta}$）と分散（$\boldsymbol{\Sigma}$）を次のような一般式で推定する。

$$\hat{\boldsymbol{\Delta}} = 1/n \times \sum_{i=1}^{n} \boldsymbol{\delta}_i$$

$$\hat{\boldsymbol{\Sigma}} = 1/(n-1) \times \sum_{i=1}^{n} (\boldsymbol{\delta}_i - \hat{\boldsymbol{\Delta}})(\boldsymbol{\delta}_i - \hat{\boldsymbol{\Delta}})^T \text{、}$$

ここで$\mathbf{n}$は遺伝子の数を表す。$\hat{\boldsymbol{\Sigma}}$に対して$\hat{P} \cdot \hat{\boldsymbol{\Sigma}} \cdot \hat{P}^T = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$が成り立つ$\hat{P}$を求め（$\mathbf{r}$は$\hat{\boldsymbol{\Sigma}}$のランクである）、真の平均値$\boldsymbol{\Delta}$の不確実性の計算に用いた。 不確実性を計算する際、遺伝子と塩基配列のサイトを再抽出する２段階のサンプリング方法を考えた。２段階のサンプリング方法で得られた$\boldsymbol{\delta}_i$の平均、分散、P 行列の推定値をそれぞれ$\hat{\boldsymbol{\Delta}}^{**}$、$\hat{\boldsymbol{\Sigma}}^{**}$と$\hat{P}^{**}$で表す。するとベイズ枠組みで$\boldsymbol{\Delta}$の分布は次のように近似できる。

$$\boldsymbol{\Delta} \sim \hat{\boldsymbol{\Delta}} - \hat{P}^{-1}\hat{P}^{**}(\hat{\boldsymbol{\Delta}}^{**} - \hat{\boldsymbol{\Delta}}) \qquad (1)$$

ある系統樹作成アルゴリズム T に対して、系統樹$\hat{T} = T(\hat{\boldsymbol{\Delta}})$の分布を調べることにより、系統樹の不確実性を定量化することができる。 この方法は遺伝子間のバリエーションを適切に考慮するという面で、従来の方法より優れているといえる。

４．研究成果

式（１）の方法を哺乳類のデータ解析に適用し、解析手法の妥当性を検証した(Seo 2008)。先行研究で発表された、10 種の哺乳類から由来する 2789 個の遺伝子を用いて、Boreotheria、Afrotheria、Xenarthra の系統関係を調べたのだ。 距離行列法を使うと、rodent グループが外群につながるという問題が先行研究で指摘されてきたが、この問題は遺伝子サイト間の進化速度の異質性を無視することに起因するということが本研究で明らかになった。「Sequence Concatenation」アプローチを使うと、遺伝子の数が多い場合、Afrotheria、Xenarthra が一緒にグループ化される系統樹が 100%のブートストラップ確率で支持される。ところが、新しい解析手法を適用すると、Boreotheria と Afrotheria がグループ化される系統樹が 42.4－80.0%の範囲で、また Afrotheria と Xenarthra が一緒にグループ化される系統樹は 15.9－57.0%の範囲で支持されるのが分かった。 この結果は 2 種類の系統樹がどちらも強く否定できないということを意味するし、最尤法

を採用した先行研究と合致することである。

複数の遺伝子を用いる研究の一環として、分岐年代の推定も行った(徐ら 2008)。69 種の哺乳類から得られた 12 種類のミトコンドリアタンパク質コード領域を解析し分岐年代と同義置換・非同義置換速度の変動を調べた。12 種類のタンパク質コード領域をつなぎ合わせず、別々に扱うことによって、同義置換・非同義置換速度の変動パタンを個別に調べることが可能になった。そのデータを解析した結果、霊長類、Afrotheria、Carnivora グループの共通祖先がそれぞれ 68.2±2.7、91.3±2.0、53.3±2.6（「平均±標準誤差」を表す）百万年前と推定されるなど数多い分類群の年代推定値が先行研究と合致し、我々の解析手法の妥当性が確認された。同義置換・非同義置換速度の変動については、哺乳類の進化の過程で強い相関を持っているのが分かり、またその原因としては集団の大きさの変動、突然変異率の変化などが考えられる。

複数の遺伝子を用いる解析の場合、計算時間の負担が大きくなるなどの問題点も予想される。そういうことを踏まえて、計算量が軽減できる配列進化モデル開発研究も行った(Seo and Kishino 2008, 2009)。新しいモデルは、アミノ酸置換モデルとコドン置換モデルの長所を融合したモデルであり、今後大量のタンパク質コード領域の解析において有効な手段になると期待している。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 7 件）

Seo T-K (2009) Classification of nucleotide sequences using Support Vector Machines. (submitted)

Seo T-K, Kishino H. (2009) Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. Syst. Biol. (in press)

Xiang, Q-Y, Thorne, J, Seo, T-K, Zhang, W., Thomas, D.,Ricklefs, R. (2008) Rates of nucleotide substitution in Cornaceae (Cornales)-correlation between molecular variation and morphological change, Mol. Phylogenet. Evol. 49:327-342.

Seo T-K, Kishino H. (2008) Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst. Biol.57:367-377

徐 泰健, 岸野 洋久, Jeffrey L. Thorne, (2008) コドンモデルを用いた分岐年代のベイズ推定, 統計数理 56:37－54

渡部 輝明, 徐 泰健, 岸野 洋久, (2008) 遺伝子型の分子進化と表現型の適応進化, 統計数理 56:55－66

Seo T-K (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960-971

〔学会発表〕（計 6 件）

Seo, T-K, Kishino, H.: Statistical Comparison of Nucleotide-, Amino-Acid-, and Codon-Substitution Models for Evolutionary Analysis of Protein-Coding Sequences, SMBE meeting (2008).

Seo, T-K, Kishino, H.: Statistical Comparison of Nucleotide-, Amino-Acid-, and Codon-Substitution Models for Evolutionary Analysis of Protein-Coding Sequences, 日本進化学会 (2008)

Seo, T-K., Kishino, H.: Synonymous substitution substantially improve evolutionary inference from highly diverged proteins. SMBE meeting (2007).

Seo, T-K, Kishino, H.: Synonymous substitution substantially improve evolutionary inference from highly diverged proteins. 日本進化学会 (2007).

Seo, T-K, Calculating bootstrap probabilities of phylogeny using multilocus sequence data. SMBE meeting (2006).

Seo T-K, Kishino H. Jeffrey L. Thorne Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. 日本進化学会 (2006)

〔図書〕(計　　件)
該当なし

〔産業財産権〕
該当なし

○出願状況（計　　件）
該当なし

○取得状況（計　件）
該当なし

〔その他〕

６．研究組織
(1)研究代表者
徐　泰健（Seo, Tae-Kun）
東京大学・大学院農学生命科学研究科・特任助教
研究者番号：６０４０１１８９

(2)研究分担者
該当なし

(3)連携研究者
該当なし