

# ベイジアン MCMC による生物間系統関係の推定法

田辺 晶史

筑波大学大学院生命環境科学研究科

## はじめに

生物間の系統関係推定法は、表 1 のような形質状態行列データから、図 1 のような系統樹を得るものである。最節約系統推定法では、候補となる系統樹においてそのデータが実現するために必要な最小置換回数を数え上げ、その値が最も小さい系統樹を最節約系統樹として採用する。しかし、塩基配列データでは、形質状態間の置換確率は異なることがあり得る。また、形質間でも置換確率は異なる

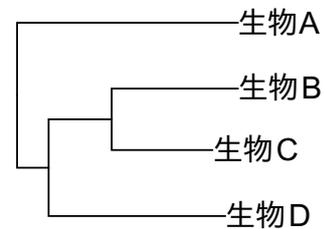
表 1 DNA 塩基配列データの例

	形質						
	1	2	3	4	5	6	7
生物 A	T	T	T	G	C	T	C
生物 B	T	T	C	C	C	T	C
生物 C	G	C	C	T	C	T	C
生物 D	C	C	T	T	C	T	C

だろう。そこで最尤系統推定法では、これらの置換確率の不均質性をモデル化して系統樹の尤度を算出する。そして、尤度を規準として候補系統樹を比較し、最も尤度が高いものを最尤系統樹として採用する。ここで注意しなくてはならないのは、最尤系統推定法は統計的には「推定」ではなく、尤度を規準とした「モデル選択」であることである。

これらの方法の最大の難点は、扱う生物の増大に伴って候補となる仮説系統樹の数が爆発的に増大することである(表 2)。そのため、数々の発見的探索法が開発されている。そのほとんどは近隣結合法 (Saitou and Nei, 1987) や段階的配列付加法 (Swofford and Begle, 1993) などにより初期系統樹を作成し、初期系統樹の樹形を少しだけ改変した周辺系統樹の尤度を計算、尤度が最高のものを採択し尤度の改善が得られなくなるまでさらにその周辺探索を繰り返すというものである。大規模データセットの解析によって

図 1 系統樹の例



比較した研究では、初期系統樹をばらつかせて生成し、それらの周辺探索を反復する方法

(Vos, 2003) と遺伝的アルゴリズムを応用した方法 (Zwickl, 2006) が優れていると言われている (Morrison, 2007)。

表2 OTU 数と想定できる樹形の数

OTU 数	想定できる 2 分岐のみの無根系統樹の樹形数
3	1
4	3
5	15
10	2,027,025
20	221,643,100,000,000,000
30	8,687,364,000,000,000,000,000,000,000,000

比較的近年になってここにベイジアン MCMC を適用しようとする研究 (Yang and Rannala, 1997; Ronquist and Huelsenbeck, 2003) が現れ、一時は主流を占めるまでになった。これは、従来はブートストラップリサンプルデータにおける解析を反復して得ていた系統樹の信頼性をも含めた結果を、大規模データセットでは大幅に短い時間で得られたためである。しかし現在では最尤系統推定を行うソフトウェアの進歩によって計算速度面での逆転を許している。

本稿ではまず簡単に系統推定の必要性と手順に触れた上で、現在のベイジアン MCMC による系統推定法を概観する。そして最尤法に対する利点と欠点を踏まえて将来のベイジアン系統推定の姿を議論する。

## 生物学における系統樹の必要性

系統樹は、表現型形質などをその上に配置 (祖先形質復元) したりすることで進化の過程を明らかにすることができる。たとえばイカやタコのカメラ眼と脊椎動物のカメラ眼が独立な異なる起源に由来することは確かな系統樹によって初めて立証される。これはコウモリと鳥の翼も同様である。これが、最も多い系統樹の使われ方である。しかし、最近になって 2 つの用途が新たに加わってきた。

そもそも、これまでに「系統的に独立した生物」は確認されていない。そのため、統計解析上の最も重要な仮定の一つである「サンプル間の独立性」が生物では成立し得ないことは明白である。ここで、仮に 100 種のトカゲの体色を考えてみる。草原と砂漠にそれぞれ 50 種生息し、草原に生息する種は緑色と褐色がそれぞれ 40 種と 10 種、砂漠に生息する種は緑色と褐色がそれぞれ 20 種と 30 種だったとしよう (表 3)。では、生息環境と体色に関連性はあるだろうか。系統関係を考慮しない研究者

表 3 トカゲの体色の例  
— 数値は種数。

	草原	砂漠	合計
緑色	40	20	60
褐色	10	30	40
合計	50	50	100

はサンプル間の非独立性には気付かず、Fisher の正確確率検定を用いてしまうかもしれない。すると、 $p = 0.00008039$  となり、生息環境と体色に関連性を見いだすことになるだろう。しかし、仮に表 3 のマス目の各種群が単系統であったならどうだろう。数えるべきは種数ではなく単系統群の数であり、各マス目の数値は 1 ということになる。以上のように、系統関係を考慮するかしないかで結論が大きく変化し得ることがわかるだろう。そこで、系統関係=サンプル間の非依存性を考慮して統計解析を行うことでこのような問題を解決しようとする手法があり、系統的独立比較法などと呼ばれている (Felsenstein, 1985; Grafen, 1989)。これが系統樹の第 2 の用途である。

系統樹の第 3 の用途は、多様性の定量化である。生物多様性の重要性が叫ばれて久しいが、いまだに多様性は種・属・科などの分類群の数で定量化されている。しかしこれらの分類群は人間が便宜的に設定したものであり、その数では十分に自然界の多様性を認識できていないと言え難いと考えられる。これを、系統樹上での生物間のパスの長さの和=系統的多様性で置き換えることで解決しようという研究が徐々に増えてきている (Faith, 1992; Forest *et al.*, 2007)。

以上のように、進化生物学に限らず個体以上のスケールを扱うマクロな生物学のあらゆる分野で、系統樹はもはや無くてはならないものであると言ってよいだろう。

## 系統推定の手順

現在の系統推定における解析手順は、細かく分けると以下の 5 段階に分けられる。

1. データ配列の整列 (相同形質の同定)
2. データ配列の区分 (同じ置換モデルを当てはめる範囲を設定)
3. 各区分領域での最適置換モデルの選択
4. 選択された置換モデルを当てはめての系統樹の選択
5. ブートストラップ法を用いた信頼性の推定 (MCMC では不要)

これにより得られた系統樹に基づいて、さらに分岐年代推定や祖先形質復元、系統的独立比較といった解析が行われる。

この手順には極めて大きな問題が潜んでいる。何故なら、配列の整列も、置換モデルの選択も、「仮の系統樹」に固定して行われるため、多重モデル選択となっているからである。さらに言えば、系統樹を用いた解析法も最適系統樹に固定してしまうため、系統樹の不確実性が十分に考慮されないことが多い。そのため、ソフトウェアから出力される不確実性は必ずと言っていいほど過小評価されているのである。

## ベイジアン MCMC による系統推定

置換モデルと系統樹が与えられれば、尤度を算出することができる。この尤度を利用して置換モデルの各パラメータ、系統樹の各枝の長さ、樹形を Metropolis-Hastings アルゴリズム (Metropolis *et al.*, 1953; Hastings, 1970) を用いて最適化していくのがベイジアン系統推定である。事前分布にはベータ分布やディリクレ分布といった共役事前分布や一様分布などの無情報事前分布が用いられる。ただし、枝長はより短い提案が多くなるような指数分布を用いることが多い。

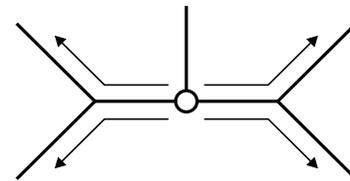
系統仮説の比較を行う際を除いて、樹形の事前分布も一様分布が用いられる。図2のような系統樹では、無作為に選ばれた内部ノードを枝に沿って無作為に移動させることで異なる樹形を提案する。そして、提案された〔置換モデル + 系統樹〕の尤度が算出され、Metropolis-Hastings のルールに従って受理または棄却される。

ベイジアン系統推定に用いられる代表的なソフトウェア MrBayes (Ronquist and Huelsenbeck, 2003) は、上述の方法を基本としつつ、それを拡張した Metropolis-coupled MCMC (MC<sup>3</sup>) (Geyer, 1991) という手法を実装している。これは、尤度の谷を越えやすくし、事後分布の広い裾野を描き出すためである。

MC<sup>3</sup> では、1つ以上の「高温系列」と実際にサンプルを得る「低温系列」を同時に走らせ、高温系列では温度に応じてより過激な提案がなされやすくする。そして任意のステップごとに全系列の中から無作為に2つの系列を選択し、高温系列側を提案と見なして Metropolis-Hastings のルールに従って〔置換モデル + 系統樹〕の状態交換を試行する。

MCMC の難点は、収束して定常状態に入っているのかという判断と、採取したサンプルが十分な量あるのかというところにある。系統推定では場合によっては数ヶ月を要することもあるため、解析を終了してから収束していなかったりサンプルサイズが不十分だったため再度行うといったことはやりづらい。そこで、MrBayes では Average Standard Deviation of Split Frequencies (ASDSF) という値を解析中に一定間隔で算出して収束を診断できるようにしている。ASDSF は、MC<sup>3</sup> を同時に2つ以上実行し、それらが定常状態に入っているのなら、推定される樹形の事後分布も一致するはずだという考えに基づくものである。複数の MC<sup>3</sup> 間で分布が一致していれば、ASDSF の値は0に近づく。一般にはこの値が0.01未満であることをもって収束したと判断している。

図2 樹形の提案方法



円で示したノードが選択された場合、矢印で示したルートが無作為に選択してノードを移動させる。

ASDSF だけでは収束しているかしかわからず、サンプルサイズが十分かどうかまではわからない。そこで、Tracer というソフトウェアによって、解析中でも有効サンプルサイズ (Kass *et al.*, 1998) を算出することができる。有効サンプルサイズが 200 以上なら十分なサンプルがあると見なされる。Tracer では、各パラメータの密度曲線やステップごとの値をプロットすることもできるため、これを見るのも非常に有効な手段である。

以上のようにして得た MCMC の解析結果から、MAP 系統樹、多数決合意樹や内分枝 = 系統仮説の出現頻度表を作成するといった要約を行うことで、ベイジアン系統推定は完了する。通常、多数決合意樹の各内分枝上には解析結果から得られた事後確率をその内分枝の信頼性の指標として示す。しかし、内分枝の事後確率はサンプル系統樹のごく一部分をサンプル横断的に取り出して算出したものである。このような事後確率の計算方法の妥当性については定かではない。ともあれ、このような図と値に基づいて生物間の系統関係と進化過程に関して議論が行われている。出現頻度表からは、最多数派ではない内分枝の事後確率も得られるため、次点の仮説を得ることもできる。

## Bayes factor による系統仮説の比較

前節の方法により、最多数派の系統仮説と次点の系統仮説の事後確率が得られるが、系統推定においては最多数派の事後確率値は過大評価になりやすいことが知られている (Suzuki *et al.*, 2002)。そこで、系統仮説を樹形制約として課した = 系統仮説を樹形の事前分布として与えた MCMC を行って結果を比較することで系統仮説の優劣をよりの確に判断できる。これには Bayes factor (Kass and Raftery, 1995) という指標の値が用いられている。この値も Tracer によって計算できる。

## 最尤系統推定法との比較

最尤系統推定 + ブートストラップ解析とベイジアン系統推定は、現状ではともに複雑な置換モデルに基づく最適系統樹 (最尤系統樹・MAP 系統樹) が得られ、系統仮説の信頼性を示す指標 (ブートストラップ支持率・事後確率) が得られる点に変わりはない。ベイジアン MCMC の際の事前分布はほとんどが共役事前分布か無情報事前分布であることもあり、ベイジアン系統推定はユーザーからは「高速な最尤 + ブートストラップ法もどき」と認識されていた。しかし、最尤系統推定のソフトウェアは速度面での改善を継続してきたため、その優位性は失われてしまった。むしろ MCMC を収束させるための各種パラメータ設定に職人技が要求されてしまう問題点の方が大きくなりつつある。ここでは「信頼性の過大評価」という短所と、「柔軟なモデリングが可能」という長所について述べる。

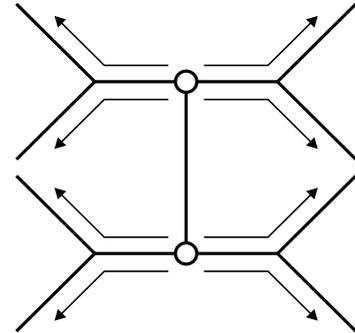
既に述べたように、事後確率は信頼性を過大評価してしまう。この理由は、おそらくは MCMC の実装と不適切な事前分布の指定にある。前述の通り MCMC では、樹形の改変はノードを枝をまたいで動かすことで実現されるが、MrBayes の場合は遠い樹形も提案されるように内分枝の両端ノードを一度に動かす (図 3)。このため、実際には 2 つのパラメータ (実は内分枝の長さも変えているため 3 つ) が動かされているにもかかわらず提案の受理確率がそのままなので不当に却下されやすくなる。これでは最多数派の事後確率が過大評価になってしまうのは当然と言える。また、枝長の事前分布に指数分布を使うと述べたが、この

事前分布によって内分枝の事後確率が大きな影響を受けることがわかっている (Yang and Rannala, 2005)。最尤系統推定法とブートストラップ解析の併用によって得られる信頼性はやや保守的な値であると言われている (Suzuki *et al.*, 2002)。ただし、これは現在主流のソフトウェアによる結果ではないため、うまく「手抜き」することで高速化している最新のソフトウェアが出力する信頼性が同様の性質を持つのかは検証の必要があるだろう。また、ベイジアン系統推定では仮説の比較を Bayes factor で行うが、この方法では多数の系統仮説を比較する際に多重性が考慮されなくなってしまう。これも最適系統樹の信頼性過大評価に繋がる。最尤系統推定とブートストラップの枠組みでは、この問題を解決する方法が開発され利用されている (Shimodaira and Hasegawa, 1999; Shimodaira, 2002)。

ベイジアン系統推定には改善の余地があることは間違い無い。しかし、最尤系統推定では不可能な柔軟なモデリングが実現できることこそベイジアン系統推定の利点である。たとえば、現在のところ系統樹の枝によって異なる置換モデルを適用することは最尤系統推定ではほとんどできない。また、置換モデルは A から C への置換と C から A への置換の確率を等しいと仮定する対称型モデルが用いられており、非対称型モデルも同様に使われていない。これは最適化が非常に難しいためである。しかし、ベイジアン MCMC によってそれを可能にした実装が存在する (Blanquart and Lartillot, 2006, 2008)。

ほとんどの場合、複数の遺伝子が同一の系統樹を支持することが仮定され解析されている。しかし、遺伝子の系統樹と種の系統樹は様々な要因により一致しない。また、複数の遺伝子間で異なる系統樹を支持することが実際に知られている (Rokas *et al.*, 2005)。そこで、遺伝子間で異なる系統樹を支持することを許容しながら種の系統樹を得る手法が求められていた。それを Liu and Pearl (2007) はベイジアン MCMC によって実現している。

図 3 MrBayes の extending-TBR



円で示した内分枝の両端にある 2 つのノードを一度に動かすことで比較的遠い樹形も提案されるようにしている。

この方法では種の実在を仮定する必要があり、各遺伝子の系統樹の違いがパラロジーや遺伝子水平伝播や収斂などによるものでなく各遺伝子の本当の系統樹の違いを反映したものであることを要するなど実用面では多くの条件をクリアする必要があるが、ベイジアン MCMC を使えばこのようなことが可能だということを示す好例である。

複雑で柔軟なモデリングを応用すれば、前に述べた系統推定の手順を減らすこともできる。系統推定の複雑な手順は不確実性を隠蔽し過小評価してしまうことになるが、手順が減ればそれだけ不確実性の考慮の漏れが減ることになる。これはおそらく最尤系統推定では実現できないだろう。既に、可逆ジャンプ MCMC によって置換モデル間の遷移を可能にした実装は存在する (Huelsenbeck *et al.*, 2004)。これによって、従来は不可能だったモデル選択の不確実性が考慮できる。また、データ配列の区分を事前知識無しに最適化しながら系統推定を行う実装もある (Pagel and Meade, 2004; Lartillot and Philippe, 2004)。データ配列の整列と系統樹の同時推定も既に実現されている (Fleissner *et al.*, 2005; Lunter *et al.*, 2005; Redelings and Suchard, 2005)。これらはいずれも前述の手順のうち、[系統樹の選択 + 他のステップ] に過ぎないが、技術的には全てのステップを統合したワンステップ系統推定がベイジアン MCMC を使えば可能であることを示している。

## 系統解析の大統合へ向けて

以上のように、ベイジアン MCMC による系統推定は極めて大きな可能性を秘めているが、全ての手順を統合するには非常に大きなハードルが存在する。それは計算資源の不足である。あまりに多くのことを考慮しようとしているため、MCMC のサンプリング間隔は非常に長くする必要があり、1 度の尤度計算だけでも計算量は激増する。それに、MCMC のアルゴリズムとしての性質が大きな足かせとなる。

現在の計算機の発展は、演算器の高速化ではなく演算器の増加とベクトル化の方向へ向かっている。演算器の高速化ではソフトウェアはそのままでも高速化した。半導体の物理化学的制約により困難となったため、演算器を増やしたりベクトル命令によって 1 度に多数のデータを処理するようになったのである。これらを生かしてソフトウェアを高速化するには、並列化とベクトル化が不可欠であるが、そもそも MCMC はその性質上、並列化には不向きである。というのも、あるステップの演算は、前のステップの演算結果に依存しており、独立性が無い極めて逐次的な処理だからである。MC<sup>3</sup> のように多数の MCMC を同時に実行すれば、各 MCMC を 1 つの演算器に割り当てる並列化によってそれなりに高速化することはできる。しかし、PC レベルならともかく PC クラスタやスーパーコンピュータの多数の演算器を生かそうとすると MCMC の数を増やす必要がある。

ところが MC<sup>3</sup> では MCMC の数を増やしてもサンプルが採取されるのはただ 1 つの低温系列であり非常に無駄が多い。状態交換試行の数も大幅に増やさなければならない。状態交換試行のためには同期を取ったりプロセス・スレッド間で通信を行う必要があるが、並列処理を律速するのはまさにそこであるため、大きな改善が必要である。MC<sup>3</sup> を多数走らせることもできるが、通信が必要なことには変わり無い。それに対して、最尤系統推定 + ブートストラップの枠組みは並列処理に非常に向いている。何故なら初期樹形から改変した候補樹形の尤度計算はそれぞれ独立しており、ブートストラップの各繰り返しも完全に独立しているためである。つまり、プロセス・スレッド間通信がほとんど必要無いのである。

1 つの方法として、MC<sup>3</sup> でないただの MCMC を演算器の数だけ実行し、実行後に得られたサンプル群から、MC<sup>3</sup> 的にサンプルを得ることはできるかもしれない。この方法では演算中は通信処理が必要無い上、事後的に MC<sup>3</sup> の数や MC<sup>3</sup> 内の系列数を決定できる利点がある。ただし、MCMC 後の処理が激増することと、収束やサンプル量が十分かといった判断が難しいことが考えられる。今後、他にも様々な手法が試みられる必要があるだろう。

データと演算処理の性質上、系統樹の尤度の算出はベクトル化に非常に向いている部分があるはずである。演算器のベクトル化や超並列ベクトル演算器とも言える GPU の普及によってベクトルプログラミングの環境も急激に変化しようとしている。既存の系統推定ソフトウェア開発者にはこれらの技術に長けた人間はほとんどいないため、技術者の新規参入が強く望まれる。

## 引用文献

- Blanquart, S. and Lartillot, N. (2006) "A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution," *Molecular Biology and Evolution*, **23**, pp. 2058–2071.
- (2008) "A site- and time-heterogeneous model of amino acid replacement," *Molecular Biology and Evolution*, **25**, pp. 842–858.
- Faith, D. P. (1992) "Conservation evaluation and phylogenetic diversity," *Biological Conservation*, **61**, pp. 1–10.
- Felsenstein, J. (1985) "Phylogenies and the comparative method," *American Naturalist*, **125**, pp. 1–15.
- Fleissner, R., Metzler, D., and von Haeseler, A. (2005) "Simultaneous statistical multiple alignment and phylogeny reconstruction," *Systematic Biology*, **54**, pp. 548–561.
- Forest, F., Grenyer, R., Rouget, M., Davies, T. J., Cowling, R. M., Faith, D. P., Balmford, A., Manning, J. C., Procheş, S., van der Bank, M., Reeves, G., Hedderon, T. A. J., and Savolainen, V. (2007) "Preserving the evolutionary potential of floras in biodiversity hotspots," *Nature*,

445, pp. 757–760.

- Geyer, C. J. (1991) "Markov chain Monte Carlo maximum likelihood," in Keramidas, E. M. ed. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Fairfax Station: Interface Foundation.
- Grafen, A. (1989) "The phylogenetic regression," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **326**, pp. 119–157.
- Hastings, W. K. (1970) "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, **57**, pp. 97–109.
- Huelsenbeck, J. P., Larget, B., and Alfaro, M. E. (2004) "Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo," *Molecular Biology and Evolution*, **21**, pp. 1123–1133.
- Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. (1998) "Markov chain Monte Carlo in practice: a roundtable discussion," *American Statistician*, **52**, pp. 93–100.
- Kass, R. E. and Raftery, A. E. (1995) "Bayes Factors," *Journal of the American Statistical Association*, **90**, pp. 773–795.
- Lartillot, N. and Philippe, H. (2004) "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process," *Molecular Biology and Evolution*, **21**, pp. 1095–1109.
- Liu, L. and Pearl, D. K. (2007) "Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions," *Systematic Biology*, **56**, pp. 504–514.
- Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., and Hein, J. (2005) "Bayesian coestimation of phylogeny and sequence alignment," *BMC Bioinformatics*, **6**, p. 83.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953) "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, **21**, pp. 1087–1092.
- Morrison, D. A. (2007) "Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences," *Systematic Biology*, **56**, pp. 988–1010.
- Pagel, M. and Meade, A. (2004) "A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data," *Systematic Biology*, **53**, pp. 571–581.
- Redelings, B. D. and Suchard, M. A. (2005) "Joint Bayesian estimation of alignment and phylogeny," *Systematic Biology*, **54**, pp. 401–418.
- Rokas, A., Krüger, D., and Carroll, S. B. (2005) "Animal evolution and the molecular signature of radiations compressed in time," *Science*, **310**, pp. 1933–1938.
- Ronquist, F. and Huelsenbeck, J. P. (2003) "MrBayes 3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, **19**, pp. 1572–1574.
- Saitou, N. and Nei, M. (1987) "The neighbor-joining method: a new method for reconstructing phylogenetics trees," *Molecular Biology and Evolution*, **4**, pp. 406–425.
- Shimodaira, H. (2002) "An approximately unbiased test of phylogenetic tree selection," *Systematic Biology*, **51**, pp. 492–508.
- Shimodaira, H. and Hasegawa, M. (1999) "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Molecular Biology and Evolution*, **16**, pp. 1114–1116.
- Suzuki, Y., Glazko, G. V., and Nei, M. (2002) "Overcredibility of molecular phylogenies ob-

- tained by Bayesian phylogenetics," *Proceedings of the National Academy of Sciences of the United States of America*, **99**, pp. 16138–16143.
- Swofford, D. L. and Begle, D. P. (1993) *PAUP: Phylogenetic Analysis Using Parsimony*, Ver.3.1. *User's Manual*: Laboratory of Molecular Systematics, Smithsonian Institution.
- Vos, R. A. (2003) "Accelerated likelihood surface exploration: the likelihood ratchet," *Systematic Biology*, **52**, pp. 368–373.
- Yang, Z. and Rannala, B. (1997) "Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method," *Molecular Biology and Evolution*, **14**, pp. 717–724.
- (2005) "Branch-length prior influences Bayesian posterior probability of phylogeny," *Systematic Biology*, **54**, pp. 455–470.
- Zwickl, D. J. (2006) "Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion," Ph.D. dissertation, The University of Texas at Austin.

連絡先: 田辺晶史

〒305-8572

茨城県つくば市天王台 1-1-1

筑波大学大学院生命環境科学研究科構造生物科学専攻