研 究 報 告 書

「ランダムグラフによるゲノム進化の確率モデリング」

研究期間: 平成 20 年 10 月～平成 24 年3月
研 究 者: 間野 修平

1，研究のねらい

ゲノムとは、個体の構成に必要最低限な遺伝情報の集合です。本研究が対象とした生命現象は、その進化、すなわち遺伝的浮動、突然変異、自然淘汰などの事象によるゲノムの時間発展です。進化研究においては再現性により仮説を実証することは困難ですから、現在の標本から過去の現象を統計学的に推測します。統計学的推測にはモデルが必要ですが、ゲノム進化には確率過程が適しています。本研究では、標本が従うランダムグラフを導出し、統計的推測における計算手法を開発しました。

2，研究成果

ゲノムは1次元に配置された点の集合ですが、各点は異なる状態をとりえます。状態とは、点が遺伝子座を指す場合には対立遺伝子、点が塩基サイトを指す場合は一塩基多型を指します。有限母集団の各点の状態を相対頻度分布により指定し、進化的事象による頻度の時間発展について、集団サイズで時間をスケールし、拡散極限をとります。これが母集団のモデルです。

統計学的推測においては、標本が従う統計モデルが必要です。尤度は頻度の函数ですが、頻度は上の拡散過程に従う確率変数ですので、階層モデルとみなします。進化研究において興味があるのは進化的事象を特徴づける超母数ですから、相対頻度を上の拡散過程が定める測度で積分して得られる周辺尤度を求めます。本研究で扱った統計学的推測は、周辺尤度を最大化する進化的事象の超母数を選ぶこと、すなわち経験ベイズです。拡散過程により定まる測度が陽に与えられている場合には容易です。無限対立遺伝子モデル（突然変異が常に新しいタイプの対立遺伝子を生じる）の抽出公式として導出された Ewens 分布でいえば、母集団のモデルは線分上の測度値拡散過程、その定常分布はPoisson–Dirichlet分布、多項抽出により得られる周辺尤度はEwens分布です。しかし、これらの分布が陽に与えられる進化モデルは例外であり、ほとんどの場合は拡散過程が定める測度や周辺尤度を閉じた形で得ることはできません。そのようなときにどのように周辺尤度を求めるかということが問題になります。

一つの遺伝子座に遺伝的浮動が働く進化モデルを考えます。母集団のモデルとしてWright–Fisher拡散$X_t(X_0=x)$、標本のモデルとして標本の系図における系統の数$Y_t(Y_0=y)$をとります。Y_tは死滅過程であり、これらのマルコフ過程は双対になります。ここで、双対とは、2つのマルコフ過程X_t，Y_tが函数$f(x,y)$について$\mathbf{E}_x[f(X_t,y)]=\mathbf{E}_y[f(x,Y_t)]$を満たすことをいいます。$f(x,y)=x^y$としますと、左辺はすべて同じ対立遺伝子が抽出される周辺尤度ですが、これは右辺の死滅過程により定まりますので、死滅過程の分布さえ得られれば、拡散過程が定める分布を求める必要はありません。このような単純な進化モデルであれば標本のモデルが系図であることは自明ですが、複雑な進化モデルについては、標本のモデルとして系図をどのようなランダムグラフに拡張するべきかがわかりません。しかし、複雑な進化モデルであっても、拡散過程の生成作用素の導出は容易です。そこで、拡散過程の生成作用素を導出し、その作用素の考察から双対として標

本のモデルを導出できると考えられます。申請当時、このようにして複雑な進化モデルの周辺尤度を求める手続きを着想しました。本研究では、この手続きをいくつかの進化モデルに適用し、標本のモデルとしてのランダムグラフを導出するとともに、統計学的推測のための計算手法を考案しました。

（1）母集団のモデリングとランダムグラフの導出

① Ancestral Bias Graph：ゲノムのある部分が別の部分に複写される現象を遺伝子変換とよびます。A,T から G,C の塩基への変換が逆よりも多いことが明らかになっています。母集団のモデルとして、複数の完全グラフからなり、各グラフが一定数のサイトからなる投票者モデルを考えました。各サイトは他のサイトの意見（A,T/G,C）をランダムに継承しますが、他のグラフのサイトから継承するときには G,C を継承することが多いとします。この拡散極限の双対として標本のモデルとしてのランダムグラフを導出し、Ancestral Bias Graph（ABG)と命名しました。ABG を生成するマルコフ過程は標本を過去に遡る相互作用する粒子系の確率モデルで、祖先の合流（coalescent）、遺伝子変換、分岐の3つの進化的事象を伴います。遺伝子変換の頻度が大きいとき、ABG は Ancestral Selection Graph（ASG)とよばれる遺伝的浮動と自然淘汰を考慮した一つの遺伝子座の進化モデルにおける標本のモデルに近づきます。この投票者モデルは分集団化した母集団において移住率が対立遺伝子のタイプに依存するモデルともみなせるので、進化的事象の偏りは自然淘汰と区別しにくいという進化学的示唆が得られました。

②Ancestral Collision Graph：グラフに値をとる粒子系と衝突：進化モデルとして、状態がグラフのノードで表される関係をもつ相互作用する粒子系を考えました。無作為に2つの粒子を選び、それらがグラフの隣り合う状態である場合に限り一方の粒子の状態を他方の粒子の状態に移します。このマルコフ過程は、粒子が存在するノードが独立集合をなすときに停止します。状態を対立遺伝子と考えれば遺伝子の組み合わせによる生殖隔離のモデル、状態を分集団と考えれば分集団化した母集団における移住のモデルとみなせます。このモデルの拡散極限の双対として標本のモデルとしてのランダムグラフを導出し、Ancestral Collision Graph（ACG)と命名しました。ACG を生成するマルコフ過程は標本を過去に遡る相互作用する粒子系の確率モデルで、衝突のみを生じます。ここで、同じ状態をもつ2つの粒子が出会い、その状態の粒子と隣接した他の状態の粒子となって分かれる事象を衝突と定義します。粒子数は一定ですので、ACG は従来考察されてきた branching coalescent 過程（①の ASG, ABG のような分岐と祖先の合流を生じる確率過程）とは本質的に異なるランダムグラフです。状態が完全グラフをなすとき、辺の有無によらない稀な状態推移を考慮しますと、周辺尤度は Ewens 分布で与えられます。任意のグラフの周辺尤度を計算することで、標本について赤池ベイズ情報量基準の意味で最もよいグラフを選択することができます。

③既に提案されていた ASG, ancestral recombination graph （ARG; 組み換えを考慮したランダムグラフ）の性質についても考察しました。ARG において、双対をなす拡散過程が定める分布のモーメントを求めることは難しいのですが、ARG の構造に着目することで閉じた形を得る手続きを与えるとともに、マルコフ連鎖モンテカルロ （MCMC） による数値計算法を考案しました。

（2）正確検定、尤度、事後分布の計算手法の開発

①正確検定における代数的手法：ゲノムのある部分の標本を並べたとき、配列のタイプの分布

は確率分割の実現です。確率分割は生命現象に限らない様々な現象に現われます。Ewens 分布は確率分割の基本的分布であり、ノンパラメトリックベイズにおける事前分布としても用いられています。最大数の大きさの分布は極値分布として興味深いので、Ewens 分布の一般化である Pitman 分布について最大数の大きさの分布の閉じた形を組み合わせ論的に導出しました。正確検定においてすべての分割を生成することは現実的ではありません。そこで、条件付き分布からの標本抽出を MCMC によりシミュレートし、数値計算しますが、分割をくまなく推移する規則が必要になります。そこで、各分割を単項式に対応させて多項式環を定義し、トーリックイデアルの Gröbner 基底に対応する推移規則を導出しました。

②周辺尤度の確率的計算手法：本研究におけるランダムグラフの構成から、周辺尤度は標本を過去に遡る相互作用する粒子系の確率モデルの母関数になりますので、各状態の周辺尤度が満たす線形微分方程式系が得られます。ただし、状態の数が非常に大きいので代数的に解くことは現実的ではありません。そこで、係数を推移確率とみなして MCMC を用います。標本を初期状態としてランダムグラフを過去に遡る向きにシミュレートし、importance sampling により推移確率を補正し、粒子数が1になったときの境界条件を perfect simulation により定めます。この手続きを ABG に適用しました。

3，今後の展開

確率分割のような様々な現象に適用できるモデリングと、それに伴う実用的計算手法を追求したいと考えています。また、喫緊の多くの課題に共通して必要とされていることですが、複雑なモデル、大規模なデータを扱えるモデリングと計算手法も追求したいと考えています。本研究とは方向性が異なり、捨象と近似が重要になりますので、本研究とは別に、すでに考察を進めています。これらのモデリング、手法をデータに適用し、様々な課題の解決に貢献していくことができればと考えています。

4，自己評価

個人型研究さきがけ「生命現象の革新モデルと展開」という数理モデルを共通言語とする領域に参加させて頂きましたので、ゲノム進化の統計的推測に必要になるモデリングと計算手法に可能な限り数理的手法を導入することを目指しました。申請時の目標は具体的で、複雑な進化モデルについて、母集団のモデルとしての拡散過程が定める分布を求めることなく周辺尤度を求める手続きを実行するというものでした。計算サーバを購入させて頂きましたので、大規模な計算を素早く行うことができ、研究を効率的に推進することができました。目標が具体的でしたので、一応達成できたと考えています。しかし、数理の深みを追求できなかったこと、幾何的手法など試したかったにもかかわらず試すことなく終了を迎えた方向性が多く残っていることは残念です。今後の展開の中で努力したいと考えています。

5，研究総括の見解

集団遺伝学において、従来のモデルのほとんどは個々の遺伝子の進化を対象としてきたが、今後は遺伝情報のセットであるゲノムの進化にも適用できる新しい数理モデルが必要になるとの問題意識のもとに、遺伝的浮動、突然変異、自然選択、遺伝子変換等の進化事象を組み入れた拡散過程の双対として、ゲノム標本の系統関係を表すランダムグラフを導出し、ゲノム進化

を統計学的に推測するという、独創的な課題に取組んだ。本人が保有する高度な統計学の知識や集団遺伝学の知識を駆使して、Ancestral Bias Graph や Ancestral Collision Graph と名付けたランダムグラフを導くなど、ゲノム時代にふさわしい新規な発想を導入したという点で高く評価できる。さらに本モデルのシミュレーションを効率的に実行するアルゴリズムの研究も同時に進められている所であり、今後は具体的な系への応用でも成果を上げてほしい。

6, 主な研究成果リスト

（1）論文（原著論文）発表

1. Shuhei Mano, “Duality, ancestral and diffusion processes in models with selection”. Theor. Popul. Biol. 75, 164–175, (2009).
2. Shuhei Mano, “Ancestral graph with bias in gene conversion”. arXiv: 0907.1127.
3. Shuhei Mano, “Duality between the two-locus Wright-Fisher Diffusion Model and the Ancestral Process with Recombination”, arXiv:1201.5557.

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物等）

Shuhei Mano, “Duality between population and sample in interacting particle systems on graphs”, *ISM Symposium on Stochastic Models and Discrete Geometry*, October 7, 2011. Tokyo.

Shuhei Mano, “Ancestral Graph with Bias in Gene Conversion”, *Molecular Evolutionary Studies in Post-Genomic Era: Workshop on Evolutionary Analyses and Applications*, June 9, 2011. Xian, China.

Shuhei Mano, “Ancestral Process with bias in ectopic gene conversion or migration”, *International Center for Mathematical Sciences Workshop: Stochastic Population Dynamics and Application in Spatial Ecology,* June 18, 2009. Edinburgh, U.K.