

## 論文審査の結果の要旨

申請者氏名 Leonardo de Oliveira Martins

ウイルスゲノムは頻繁に組換えを経験する。このため、ゲノム配列の多型性は単一の系統樹では表現できず、領域により異なる系統関係（トポロジー）を持つ。その結果、ゲノムの組換えを無視して解析を行うと、ウイルスゲノムの分子進化について、誤った推論をしてしまう危険性がある。ウイルスゲノムの組換えを推定する方法として、大きく分けて二通りのアプローチがある。ひとつは集団遺伝学的なアプローチである。これはゲノム上の部位の間の連鎖不平衡の情報を下に、組換えの履歴を祖先組換えグラフ(Ancestral Recombination Graph)で表現し、組換え率を推定する。ただし、連鎖不平衡の強さは集団の履歴やゲノムにかかる淘汰圧に影響される。しかし、これらのモデル表現は複雑であることから、現在のところ、組換えの推定プログラムは中立進化と平衡集団を仮定している。

もう一つは分子系統学的なアプローチである。この方法は、ゲノムの領域の間でトポロジーが食い違うことを利用して組換えを検出するもので、淘汰圧や集団の履歴に関する仮定を置く必要がない。比較的ゲノム組換えの頻度が小さく、組換え位置に挟まれたゲノム断片が分子系統に係るシグナルを保持しているときに有効である。組換え親が予め知られているときには、彼らの配列を参照配列として、対象とする配列を領域に分割し、参照配列に関連付ける。検出力を高めるために、トポロジーを状態とする隠れマルコフモデル、組換え位置に関するベイズ型変化点モデルが開発されてきている。組換え親が未知の場合は、すべてのトポロジーの可能性を調査する必要がある。その自由度の高さ（配列数が 10 本ですでに可能なトポロジーの数は 200 万を超える）から、扱える配列の数は数本と限定されるのが弱点である。本研究では、組換え距離に関する事前分布を導入することにより、この弱点を克服した。

### 1. SPR 距離の近似アルゴリズム

本論文で提案する方法は、隣接領域間におけるトポロジーの食い違いのうち、組換えによる成分にペナルティを課す。すなわち、トポロジーの推定に伴う誤差は不規則であるのに対し、組換えによるトポロジーの食い違いは規則性を持っていることに注意する。トポロジー空間における実質的な自由度を低く抑えることにより、計算負荷の重圧から解放されるとともに、組換えの検出力が格段に向上することが期待される。トポロジーの間の距離としては、Robinson-Foulds 距離と補最大一致部分樹(complementary maximum agreement subtree: cMAST)距離が知られている。前者は対応関係のない枝の数で定義され、後者は極大共通部分樹の葉の補集合で定義される。残念ながら、これらの距離はいずれも、組換えの回数との関連が薄い。組換えによるトポロジーの食い違いは、部分樹刈り込み・すげ替え(subtree prune-and-regraft: SPR)距離と関係してい

る。本論文では、トポロジー対の縮約表現と枝の抜き取りの最節約的更新アルゴリズムを開発し、トポロジー間の食い違いを説明する必要組換え回数を近似計算することが可能となった。

## 2. 組換えのベイズ推定法

分子進化をマルコフ過程でモデリングすることにより、各サイトの尤度が記述できる。この尤度は、トポロジーと隣接ノード間の推移確率で表現される。分子進化速度のサイト間の不均質性、サイト間で加速・減速する枝が異なる heterotachy に対して頑健な推定をするために、枝長にランダム性を導入し、確率変数とした。隣接したサイトのトポロジー不一致に対する SPR 距離にポアソン分布に従う事前分布を導入することにより、ベイズの枠組みで組換え頻度に対するペナルティを実現した。各サイトが個別に、異なるトポロジーを持つことを許す。組換えに対するペナルティの強さ、およびサイト間の不均質性は、予め固定することなく、事前分布を規定する超パラメータにランダム性を導入し、階層ベイズモデルを構築した。マルコフ連鎖モンテカルロ法 (MCMC) により、超パラメータとサイトごとのトポロジー、隣接サイト間の SPR 距離の事後分布を求める。これにより、組換え位置と組換えのパターンを同時推定することが可能となる。

## 3. シミュレーションによる有効性評価と南アメリカ HIV-1 集団の解析

シミュレーションを通じて、開発ソフトウェアの有効性の検討を行うと共に、南アメリカ HIV-1 集団の解析を行った。シミュレーションでは、HIV-1 ゲノムの解析を想定し、10 サイトを1つの単位として解析することを念頭に置いて塩基置換速度を設定した。配列数 8, 12, 16 の解析を行ったところ、いずれにおいても、配列内に変異性の高い部位が存在しても偽陽性を拾うことなく、組換え位置と組換えパターンを偏りなく推定することが示された。さらに、分子系統樹のベイズ推定を行う MrBayes との対比を行った。このプログラムは組換え位置の推定は行わないため、MrBayes では組換え位置を既知として、それらに挟まれた領域のトポロジーを推定した。驚くべきことに、本研究の手法は組換え位置は未知としているにも拘らず、推定精度は高まった。これは、隣接領域間の食い違いに対する事前分布の有効性を示している。南アメリカ HIV-1 集団からサンプリングされた BF 組換え型 16 本のゲノムの解析からは、 $37 \pm 7$  の組換え位置、 $65 \pm 10$  回の組換えが検出された。それらの多くは、サブタイプ内の組換えであるため、組換え親を所与にした解析では検出されないことが確認された。

ゲノムの組換えは、適応進化において重要な役割を担うと考えられながらも、これまで推定が困難であった。本論文では、隣接領域間の SPR 距離に事前分布を導入した階層ベイズモデルを開発し、その有効性を確認した。また南アメリカ HIV-1 集団の解析からはゲノム全体にわたり数多くの組換えが検出され、これまで蓄積された HIV-1 集団に関する解析に再検討を促すこととなった。本論文で提案したゲノム組換えの頑健推定法は、数多くの進化研究者のニーズに応えるものと期待される。現在のところ、実用的な時間内で解析可能な配列の数は数十本と限られるが、今後こうした計算上の制約からも解放されるであろう。したがって、ここで得られた成果は、学問的にも応用的にも貢献するところが大きい。よって審査委員会委員一同は本論文が博士（農学）の学位を受けるに十分な価値があると認めた。