氏　　名	徐　泰健
学位（専攻分野）	博士(学術)
学位記番号	総研大甲第623号
学位授与の日付	平成１４年３月２２日
学位授与の要件	先導科学研究科　生命体科学専攻 学位規則第４条第１項該当
学位論文題目	Statistical Analysis of Viral Sequences : Bridging sampling design, molecular phylogenetics and population genetics
論文審査委員	主　査　助教授　橋本　哲男 教授　長谷川　政美 助教授　颯田　葉子 教授　岸野　洋久（東京大学）

論文内容の要旨

The high pace of viral sequence change means that variation in the times at which sequences are sampled can have a profound effect both on the ability to detect trends over time in evolutionary rates and on the power to reject the molecular clock hypothesis. Trends in viral evolutionary rates are of particular interest because their detection may allow connections to be established between a patient's treatment or condition and the process of evolution. Variation in sequence isolation times also impacts the uncertainty associated with estimates of divergence times and evolutionary rates. Variation in isolation times can be intentionally adjusted to increase the power of hypothesis tests and to reduce the uncertainty of evolutionary parameter estimates, but this fact has received little previous attention. I provide approximations for the power to reject the molecular clock hypothesis when the alternative is that rates change in a linear fashion over time and when the alternative is that rates differ randomly among branches.

When the evolutionary rate changes linearly, it can be shown as $r(t) = a(t - t_1) + r$ where t is current time, t_1 is the time of origin and a is the amount of increase or decrease per unit time. For given a, we can calculate the power to reject the null hypothesis(H_0 : $a = 0$) using the fact that the statistic $2\Delta \log L = 2 \log \frac{L(\mathbf{X}|\hat{r},\hat{a},\hat{t})}{L(\mathbf{X}|\hat{\hat{r}},0,\hat{\hat{t}})}$ tends to a non-central χ^2 distribution under alternative hypothesis ($H_1 : a \neq 0$) where the single circumflex ($\hat{}$) and double circumflex ($\hat{\hat{}}$) respectively denote maximum likelihood estimators (m.l.e.'s) under H_1 and H_0.

When the rates differ randomly among branches, we can consider the gamma distribution as a model of rate variation. If we further assume the number of substitution in each branch follows Poisson distribution, the probability density function of the number of substitutions is that of negative binomial distribution. The power to reject the null hypothesis(H_0: Evolutionary rate does not vary) can be calculated using non-central χ^2 distribution.

When the evolutionary rate is constant, the standard deviation of estimated evolutionary rates and divergence times can be approximated using Fisher information matrix. I illustrate how these approximations can be exploited to determine which viral sample should be sequenced when samples representing different dates are available.

Using pseudo-maximum likelihood approaches to phylogenetic inference and coalescent theory, I develop a computationally tractable method of estimating effective population size from serially sampled viral data. In this method, a two stage estimation procedure is adopted. The vector of times of internal nodes ($\tilde{\mathbf{t}}$) is estimated from sequence data and then these estimated node times serve as the basis for inferring effective population size ($\widehat{N_e}(\tilde{\mathbf{t}})$). Because the main interest is effective population size and not times of internal nodes, the

internal node times are nuisance parameters in my analysis and the number of these nuisance parameters increases as the number of sequences increases.

The variance of the maximum likelihood estimator of effective population size is approximated as

$$Var_{\mathbf{t},\mathbf{x}}(\widehat{N}_e(\tilde{\mathbf{t}})) \simeq \frac{N_e^2}{n-1} + \frac{1}{2^2(n-1)^2} E_{\mathbf{t}} \left\{ Var_{\mathbf{x}} \left(\sum_{i=n}^{2} i(i-1)\tilde{t}_i \middle| \mathbf{t} \right) \right\}$$

where n is the number of sequences.

I show that the variance of the maximum likelihood estimator of effective population size depends on the serial sampling design only because internal node times on a coalescent genealogy can be better estimated with some designs than with others. Given the internal node times and the number of sequences sampled, the variance of the maximum likelihood estimator is independent of the serial sampling design.

I estimate the effective size of the HIV-1 population within nine hosts. If I assume that the mutation rate is 2.5×10^{-5} substitutions per generation and is the same in all patients, estimated generation lengths vary from 0.73 to 2.43 days per generation and the mean (1.47) is similar to the generation lengths estimated by other researchers. If I assume that generation length is 1.47 days and is the same in all patients, mutation rate estimates vary from 1.52×10^{-5} to 5.02×10^{-5}. The results indicate that effective viral population size and evolutionary rate per year are negatively correlated among HIV-1 patients.

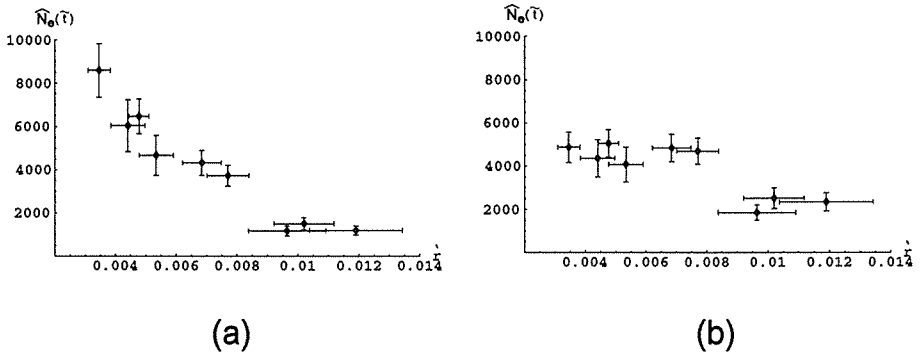

Figure 1: A negative correlation between the evolutionary rate per year $\hat{r}$ and the effective population size $\widehat{N}_e(\tilde{\mathbf{t}})$. (a) assuming a generation length of 1.47 days, (b) assuming a mutation rate of 2.5×10^{-5} substitutions per generation.

論文の審査結果の要旨

本論文の主要な貢献は以下の点である。

(1)分子時計仮説「進化速度一定」を帰無仮説としそれを棄却する検定において、対立仮説が「進化速度が線型的に変化」および「進化速度がランダムに変化」の2つの場合についてそれぞれの検出力の近似式を導くとともに、進化速度や分岐時間の推定値の誤差を評価する近似式も求めた。さらに、経時的な血液・組織サンプルが多数存在する場合に、それらをどのようにサンプリングして配列データを得ればこれらの推測をより正確に行えるのか、という具体的な実験計画の問題について議論した。

(2) 擬最尤法によるアプローチを系統樹の推論とCoalescent理論に適用し、経時的な配列データからウイルス集団の有効な大きさ(Ne)およびその分散の推定値を与える近似式を導いた。さらに提案した方法を、最近別の研究者によって提案されたベイズ的アプローチに基づく方法と比較して議論した。

(3) 9人のエイズ患者のHIV-1について得られているenv遺伝子のC2-V5領域の経時的配列データを用いて、各患者体内でのHIV-1集団の有効な大きさとHIV-1の進化速度を推定し、集団の有効な大きさと進化速度が患者間で負の相関関係にあることを明らかにした。さらに、この知見を集団遺伝学的な視点から議論した。

本論文では、進化速度が極めて速く、配列の経時的な変化が追跡可能なレトロウイルスの進化を扱う場合を想定し、経時的配列データから分子進化学的・集団遺伝学的推論を行う際の統計解析法について論じ、いくつかの新たな解析手法を提案している。既存の方法論的研究の多くが、現時点での配列データ(contemporaneous data)の解析のみを想定したものであったのに対し、レトロウイルス配列データの特殊性に注目し、新しい発想のもとで新たな方法を開発した点は評価される。また、提案した方法を具体的な実験計画の問題に適用するために、シミュレーション研究を行ったり、提案した方法に基づく実データの解析から新知見を得たことも評価される。とくにエイズウイルスの研究に関しては、今後、本論文で確立された方法に基づくさまざまな解析結果が、他の臨床的・病理的所見のデータとリンクされ、ウイルスの宿主適応の過程の解明に大きく貢献しうるものとなることが期待される。

以上の評価より、本論文の内容は博士(学術)に十分に値するものであると判定した。なお、本学位申請論文の内容に関する2編の原著論文が、国際学術誌である*Bioinformatics*と*Genetics*に掲載予定となっている。

先導科学研究科における課程博士の授与に係わる論文審査等に関する規定に基づき、公開の論文発表会を開催した。研究内容は分かり易くまとめられ、質疑応答も適切なものであった。その後審査委員による口頭試問を行った。その結果、申請者は、統計的データ解析の方法論の開発に精通していると同時に、具体的な生物学的問題を解析する能力にも秀でていると認められた。さらに、申請者が生命体科学及び関連する分野に関して十分な学識を有することも認められた。また、論文の口頭発表および論文の作成ともに英語が明快であることから判断して、申請者が十分な英語力を有することも認められた。以上の結果から試験に合格したと判定した。