氏　　　　名	Yang Zhong		
学位（専攻分野）	博士（理学）		
学 位 記 番 号	総研大乙第 162 号		
学位授与の日付	平成１８年３月２４日		
学位授与の要件	学位規則第６条第２項該当		
学 位 論 文 題 目	Bioinformatics for the study of biodiversity		
論 文 審 査 委 員	主　　査	教授	池村　淑道
		教授	長谷川　政美
		助教授	足立　淳
		教授	舘野　義男

論文内容の要旨

This thesis is a study of phylogenetic approaches and database system as well as their uses in bioinformatics. It focuses on three main topics: (A) molecular phylogenetic analysis as an effective tool to investigate the evolutionary relationships and rates and adaptation of two important groups: the mangrove family Rhizophoraceae and the sere acute respiratory syndrome (SARS) coronavirus; (B) statistical model and computer simulation approach for testing hybridization hypotheses based on incongruent gene trees; and (C) a new data model and comparison method for interacting classifications and phylogenetic trees in a taxonomic database.

In Chapter 1, I outlined the advances in today's biodiversity science and bioinformatics, and in the studies of molecular evolution and phylogenetics. To meet the major needs for a newly formed cross-disciplinary between biodiversity science and bioinformatics, *i.e.*, biodiversity informatics, applications of phylogenetic approaches and data models as well as taxonomic database systems in this field are needed.

In Chapter 2, I investigated the phylogenetic relationships and evolutionary rate heterogeneity of the family Rhizophoraceae based on the sequences of chloroplast genes *mat*K and *rbc*L, and ITS regions of nuclear ribosomal DNA. Phylogenetic trees were constructed using the maximum parsimony (MP), neighbor-joining (NJ) and maximum likelihood (ML) methods. The partition-homogeneity tests indicated that the data sets were homogeneous, and the combined analysis showed that four mangrove genera formed a monophyletic group and the terrestrial genus *Pellacalyx* was shown to be the basal clade. Evolutionary rate heterogeneity for the plastid *mat*K and *rbc*L genes in different species of the Rhizophoraceae was analyzed by means of the relative-rate tests. A number of significant rate differences at synonymous and non-synonymous sites were detected in the two genes. Two significant contrasts are that the mangrove genus *Bruguiera* has relatively slower substitution rates than the terrestrial genus *Carallia* at both synonymous and non-synonymous sites in the *mat*K sequences. The Mantel tests showed that the synonymous and non-synonymous relative-rate matrices are correlated at the *mat*K gene, suggesting that selective constraint at non-synonymous

sites is fairly constant among evolutionary lineages of the *mat*K locus. Second, there are 13 significant contrasts at non-synonymous sites in the *rbc*L sequences. Among them, six indicate that the mangrove genera have relatively faster non-synonymous substitution rates than the related terrestrial groups. However, the terrestrial genus *Carallia* still shows a relatively faster non-synonymous rate than the mangrove genus *Kandelia*. Moreover, the *rbc*L non-synonymous sites also exhibit rate heterogeneity among the terrestrial groups, regardless of their geographical distributions. The Mantel tests show that the *rbc*L rates at synonymous and non-synonymous sites are uncorrelated. The molecular evolutionary pattern of mangroves and their terrestrial relatives in which non-synonymous and synonymous substitution rates are uncoupled suggests that selection is probably an important influence on the rate variation.

In Chapter 3, I detected the adaptive evolution in SARS coronavirus (SARS-CoV) genome. First, 61 SARS coronavirus (SARS-CoV) genomic sequences derived from the early, middle, and late phases of the SARS epidemic were analyzed together with two viral sequences from palm civets. The neutral mutation rate of the viral genome was constant but the amino acid substitution rate of the coding sequences slowed during the course of the epidemic. Between the sequences of the palm civets and each of the human SARS-CoV sequences, the ratios of the rates of nonsynonymous to synonymous changes (K_A / K_S) for the S gene sequences were always greater than 1, indicating an overall positive selection pressure. However, pairwise analysis of the K_A / K_S for the genotypes in each epidemic group shows that the average K_A / K_S for the early phase was significantly larger than that for the middle phase, which in turn was significantly larger than the ratio for the late phase, which in fact was significantly less than 1. These data indicated that the S gene showed the strongest initial responses to positive selection pressures, followed by subsequent purifying selection and eventual stabilization. Second, I further tested the hypothesis that radical amino acid replacements in the spike protein, favored by environmental selective pressure during the process of SARS-CoV interspecific transmission. I investigated 108 complete sequences of the SARS-CoV S gene, and reconstructed the most recent common ancestor (MRCA) sequences of the S gene and detected the adaptive evolution in the spike protein. The results showed the simultaneous amino acid replacements in three sites, *i.e.*, 360, 665 and 701. These sites led to the excess of observed radical substitution number over corresponding expectation under the assumption of selective neutrality, indicative of potentially important roles they played in the adaptive evolution of the spike protein.

In Chapter 4, I characterized certain distinctions between hybridization and other biological processes, including lineage sorting, paralogy, and lateral gene transfer, that are responsible for topological incongruence between gene trees. Consider two incongruent gene trees with three taxa, A, B, and C, where B is a sister group of A on gene tree 1 but a sister group of C on gene tree 2. With a theoretical model based on the molecular clock, we demonstrated that time of divergence of each gene between taxa A and C is nearly equal in the case of hybridization (B is a hybrid) or lateral gene transfer, but differs significantly in the case of lineage sorting or paralogy. After developing a bootstrap test to test these alternative hypotheses, we extended the model and test to account for incongruent gene trees with numerous taxa. Computer simulation studies supported the validity of the theoretical model and bootstrap test when each gene evolved at a constant rate. The computer simulation also suggested that the model remained valid as long as the rate heterogeneity was occurring proportionally in the same taxa for both genes.

Finally, in Chapter 5, I described an information-theoretic view, *i.e.*, taxon-view, which can be applied to biological classification to capture taxonomic concepts as data entities and to develop a system for managing these concepts and the lineage relationships among them. A new data model and methodology for comparing interacting classifications were outlined. On the basis of the data model and comparison and query methods, a prototype taxonomic database system called HICLAS (Hierarchical CLAssification System) was built to query classification data and to compare interacting classifications and phylogenetic trees.

（論文審査結果）

本博士論文では、「生物多様性研究のためのバイオインフォーマティックス」というタイトルで、分子系統学的解析と分類データベースの構築について詳しく述べられている。分子系統学は、生物多様性を理解する上での基本であるが、本論文ではマングローブ Rhizophoraceae と SARS コロナウイルスについての分子系統学的な解析が行われた。第１章序論に続いて、第２章で、海水適応した種子植物であるマングローブを含む科である Rhizophoraceae について、葉緑体の２つの遺伝子 *mat*K と *rbc*L を解析し、特異な適応を遂げた Rhizophoraceae 内のマングローブ種が単系統のグループであることを示した。さらに、分子進化速度の系統間での変化を解析し、マングローブ種が陸上の近縁種にくらべてアミノ酸置換速度が速くなっており、この系統で適応的な分子進化が起っていた可能性を示唆した。

第３章では、SARS コロナウイルス 61 株のゲノム配列データをハクビシンから分離された２つのウイルス配列と併せて解析し、S 遺伝子において非同義置換速度/同義置換速度の比が１よりも大きく、さらにこの比が流行の初期において特に高くなっており、流行の後期にはこの比が小さくなっていることが明らかになった。つまり、このウイルスがはじめてヒトに感染するようになった時期において、特に急速な適応的な進化が起ったことが示された。

第４章では、遺伝子ごとの系統樹間の不一致の原因について、植物の進化において重要な雑種形成と、その他の生物学的要因、例えば系統ソーティング、パラロジー、遺伝子水平移動などとを区別するための検定法を開発した。

第５章では、これまでの生物分類データベース（ＤＢ）は分類体系が固定した静的なＤＢだったのに対して、分類体系の構造の変化に対応できる動的なＤＢのモデルを考案し、新しいＤＢを開発した。このＤＢは、分子系統学の発展により常に更新されつつある系統関係を分類体系に動的に反映できる点が優れている。

以上のように、本博士論文には多数の重要な新知見が盛り込まれており、本論文は博士（理学）に十分値するものであると判断した。なお、本論文の主要部分は、申請者を筆頭著者として、いくつかの国際学術誌に掲載されている。