論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Speedy double bootstrap method and its application for assessing the statistical reliability of estimated phylogenetic trees
著者(和文)	任愛珍
Author(English)	aizhen ren
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第9266号, 授与年月日:2013年9月25日, 学位の種別:課程博士, 審査員:渡辺　治,秋山　泰,間瀬　茂,三好　直人,杉山　将
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第9266号, Conferred date:2013/9/25, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	数理・計算科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（理学　）
学籍番号： Student ID Number		指導教員（主）： Academic Advisor(main)	渡辺治
学生氏名： Student's Name	任愛珍	指導教員（副）： Academic Advisor(sub)	秋山泰

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

This thesis entitled "Speedy double bootstrap method and its application for assessing statistical reliability of phylogenetic trees" and consists of six chapters that are relatively dependent.

In Chapter 1 "Introduction", we state the background, problem statement and our contributions. Evaluating the reliability of phylogenetic trees is critically important in the field of molecular phylogenetics and for other endeavors that depend on accurate phylogenetic reconstruction. The goal is to find an accurate and fast way of evaluating the reliability of phylogenetic trees. The bootstrap method is a well-known computational approach for assessing the reliability of phylogenetic trees. However, it is known to be biased under certain circumstances, calling into question its accuracy. Therefore, several advanced bootstrap methods have been developed to achieve higher accuracy, one of which is the double bootstrap method (DBP). This method has 3rd order accuracy. However, its complexity is very large. And another method is multiscale bootstrap method (AU). It also has 3rd order accuracy and has less complexity than double bootstrap method. This method has been shown successful in many real world applications. However, the problem of multiscale bootstrap method is that the complexity is still large.

In Chapter 2 "Speedy double bootstrap method for assessing the reliability of phylogenetic trees", we state our primary contribution, namely speedy double bootstrap method for assessing the reliability of phylogenetic trees. In this chapter, we use the PAVA (Pool Adjacent Violators Algorithm) to calculate the log-likelihood vector's projection to the boundary of each hypothesis. And we also proposed signed distance using log-likelihood vector. Through these endeavors, we proposed speedy double bootstrap method (sDBP). In theoretically, it also has 3rd order accuracy and has less complexity than double bootstrap method as well as multiscale bootstrap method and is the most close to the BP-method. For comparison, we also present the DBP-method for assessing the statistical reliability of phylogenetic trees. Finally, we discuss the advantages and disadvantages of the proposed methods.

In Chapter 3 "Evaluation of speedy double bootstrap method using biological data", we evaluate sDBP, DBP, AU and BP-method using biological data. we analyze the mammalian mitochondrial protein sequences of 6 species and the mammalian mitochondrial amino acid sequences and 12S and 16S rRNA genes of 20 species. In addition, we use our experiment results to compare the sDBP and DBP-values using the paired t-test. Through our experiment result, providing no evidence of a significant difference between sDBP and DBP-methods. We also investigate the time taken to calculate a p-value for a single tree, we compare 4 methods: DBP, sDBP, AU and BP. We conducted two separate sets of analyses. Through our experiment result, we can see that between, sDBP and DBP, sDBP is much faster. Between sDBP and AU, sDBP is faster , and got much faster when bootstrap resampling number was big.

In Chapter 4 "Evaluation of the rejection probabilities for four bootstrap methods", our focus switches to simulations and data analysis. We use simulations based on artificial data to investigate the relative statistical performance (sDBP, DBP, AU and BP). We design and perform a simulation from the normal model, and use our simulation results to compare the rejection probabilities for sDBP, DBP, AU and BP-method. Graphically, the rejection probability of sDBP, DBP and AU-method are similar, whereas that of BP-method is noticeably different. This study is the first to systematically compare these competing bootstrap probability methods via simulations.

In Chapter 5 "Implementation of speedy double bootstrap method for phylogenetic trees", we develop an easy-to-use R package, named SDBP for assessing the reliability of phylogenetic trees based on sDBP-method. We explain the implementation of our package and describe its usage when applied to the mammalian mitochondrial acid sequences and 12S rRNA and 16S rRNA genes for 20 species. We hope SDBP will be further utilized to assess the reliability of phylogenetic trees. In addition, our implementation of sDBP-test does not involve difficult calculations, such as the optimization of non-linear functions necessary for the AU-method. Therefor this method could be easily incorporated into any of the general phylogenetic analysis packages that calculate site-wise log-likelihoods from the dataset and model selection packages that

use the maximum-likelihood criterion.

Finally, in Chapter 6 “Conclusion”, we summarize and discuss the results from Chapters 2 to 5 and provide an outlook on possible future directions of research based on the methods and results presented here. Our calculations show that the sDBP-test is not confined to general tree selection problems, as the algorithm and theory can be used across various fields. Therefore, our work opens the door to many practical model selection problems that use the maximum-likelihood criterion. The procedure can be recognized as similar to sDBP-test in tree selection problems. However, our sDBP-test cannot be adapted to assess the reliability of individual nodes in a phylogenetic tree. For such individual nodes, the problem is that the PAVA method cannot be applied and the signed distance is not well defined. The same problem occurred in the assessment of uncertainties in hierarchical cluster analysis.

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 2 部提出してください。

Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 2 copies of 800 Words (English).