氏　　名	池端 久貴
学位(専攻分野)	博士(統計科学)
学 位 記 番 号	総研大甲第 1918 号
学位授与の日付	平成29年3月24日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Bayesian inference using advanced Monte Carlo methods in bioinformatics and cheminformatics
論 文 審 査 委 員	主　査　教授 福水 健次 准教授 吉田 亮 准教授 持橋 大地 准教授 山西 芳裕 九州大学

(別紙様式 2)
(Separate Form 2)

論文の要旨
Summary (Abstract) of doctoral thesis contents

This thesis describes how to approach problems on specific scientific applications with the Bayesian methods. The Bayesian methods are potentially useful to solve the inverse problem in various fields including science and industry, but real problems are not typically simple. To understand those complicated systems, a model also needs to be complicated as having many types of unknown interacting parameters. From these reasons, inferring becomes complicated not to be achieved only by a conjugate prior or a standard Monte Carlo inference. When dealing with a problem involving the high dimensional parameter space, simple Monte Carlo methods such as the rejection sampling or importance sampling are numerically infeasible due to the curse of dimensionality. Although the Markov chain Monte Carlo (MCMC) is often used as an alternative way, it has a serious drawback typically known as a local-trap problem. To deal with the local-trap problem, many existing methods use a tempering technique to make lower energy barrier between two different modes. We developed the new MCMC method, called the repulsive parallel MCMC (RPMCMC). It generates parallel Markov chains, and use repulsive forces among each chain to explore entire sampling space. A few methods including RPMCMC were confirmed to work well for a synthetic multi-modal target distribution by comparing to a simple Metropolis sampler.

As a main contribution of this thesis, two novel applications based on the Bayesian modeling were introduced. The first problem we dealt in this thesis is considered important in bioinformatics, which is called the motif discovery problem. The goal of this problem is to find recurring patterns of conserved short strings that appear in a large fraction of nucleotide sequences. It leads to discover important biological process since these recurring patterns have been preserved so that they can possibly be responsible for one of its process. Since recent experimental technologies called ChIP-seq produces much more number of fractions than before, many existing old-style algorithm need to be reconstructed to deal with large data within an acceptable time. One critical drawback of current methods such as the standard Gibbs sampling, as with the EM algorithm, arises from the following fact: the posterior distribution can be considered highly multimodal because many diverse motifs are possibly present in given sequences. Once the generated Markov chain is stuck in a locally high probability region, it is difficult to escape from that region within a finite time. This problem has received little attention in previous studies. The aim of developing our proposed method based on the RPMCMC is to achieve high detection accuracy while keeping the computational efficiency at an acceptable level. The proposed method is designed to detect many diverse motifs that previous

methods are unable to discover. In experiment, compared to the original method using a standard Gibbs sampler, this all-at-once interacting parallel run can detect much more diverse motifs. Furthermore, this method was comprehensively tested on synthetic promoter sequences and real ChIP-seq datasets. In the synthetic promoter analysis, RPMCMC found around 1.5 times as many embedded motifs as existing methods did. For the ChIP-seq datasets, the RPMCMC algorithm reported much more reliable cofactors in total than the recently published ChIP-tailored algorithms. The second problem aims at finding new molecules having some beneficial properties in a statistical manner. Computational molecular design has a great potential to save enormous time and cost in the discovery and development of functional molecules. The objective is to computationally create new promising molecules that exhibit various kinds of desired properties. Some previous studies tackled this issue with genetic algorithms (GAs) and molecular graph enumeration. The primal problem of these methods, generating unfavorable structures through their process, was avoided by introducing many incomprehensive rules. An alternative called fragment assembly method suffers from restricted design space and large computational loads. Our Bayesian molecular design begins by obtaining a set of machine learning models that forwardly predict properties of a given molecule for multiple design objectives. These forward models are inverted to the backward model through Bayes' law, combined with a certain prior distribution. This gives a posterior probability distribution conditioned on a desired property region. Exploring high-probability regions of the posterior with the sequential Monte Carlo (SMC) method, molecules that exhibit the desired properties are computationally created. The most distinguished feature of this workflow lies in the backward prediction algorithm. In this study, a molecule is described by a ASCII string, which is called SMILES format. To reduce the occurrence of chemically unfavorable structures, a chemical language model is trained, which acquires commonly occurring patterns of chemical substructures by the natural language processing of the SMILES language of existing compounds. The trained model is used in the SMC algorithm to recursively refine SMILES strings of seed molecules such that the properties of the resulting molecules fall in the desired property region while eliminating the creation of unfavorable chemical structures. The method was demonstrated with case studies in multi-objective molecular design aimed at the physical properties (HOMO-LUMO gap and internal energy) and bio-activities for 10 target proteins.

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨

Summary of the results of the doctoral thesis screening

本論文の前半部分で論じられているモチーフ発見問題は生物情報学の古典的問題であり、これまでに数多くの解析手法が提案されてきた。ところが、次世代シーケンサ（NGS）の登場によって解析対象のデータサイズが大規模化したことで、従来の方法が計算量と検出力の両面において機能しなくなり、ポスト NGS の開発競争が始まることになった。しかしながら、新世代アルゴリズムの設計原理では、計算速度の改善を優先する一方で検出力については比較的軽視されてきた。RPMCMC は、極めてシンプルな発想で従来法の問題点を克服している。論文では、他の手法との性能比較において圧倒的な検出力を有することが示されている。生物情報学における学術的貢献度は十分な水準に達しているといえる。

本論文の後半で論じられている物質探索の研究では、これまではフラグメント法というテクニックが広く適用されてきた。既存化合物の部分構造ライブラリを構造改変の部品として用いることで、化合物ライクな構造を発生するという発想である。しかしながら、生成可能な物質はライブラリのサイズの制限を受け、さらに構造置換に要するグラフ演算の計算負荷が問題視されてきた。自然言語処理の技法を用いることで、これらの問題を一気に解決するという発想は非常に先進的であり、高く評価される。

これらの研究は、生物情報学と化学情報学において重要な学術的貢献を果たすと同時に、統計科学の観点からも学術的価値が十分に認められる。モチーフ解析の論文は、Bioinformatics 誌に掲載されている（査読付き論文、第一著者)。物質探索の論文は、Journal of Computer-Aided Molecular Design 誌（査読付き論文、第一著者）に採録が決まっている。以上より、博士論文審査委員会は全員一致で学位出願論文が博士（統計科学）の学位授与に十分に値するものと判定する。

以上の論文の評価に加え，口述による試験を実施した結果、出願者はその博士論文を中心としてそれに関連がある専門分野､その基礎となる分野及び英語の能力について博士(統計科学）の学位の授与に十分な学識を有するものと判断し、合格と判定した。