別紙２

Results report

１．Title of Research and Development： Privacy-preserving genomic data analysis for personalized medicine (PRIVAGEN)

２．Principal Investigator：Kana Shimizu （Senior Research Scientist, National Institute of Advanced Industrial Science and Technology）

３．Counterpart Principal investigator：Antti Honkela（Academy Research Fellow（Finland））

４．Results of Research and Development:

In this fiscal year, the Japanese team has focused on developing cryptographic protocols for searching genome information.

1) Privacy-preserving allele information search system

We implemented a novel system in which the user can search allele information without showing his/her query to the server while the server only returns the search result. To achieve practical performance, we developed an improved protocol whose communication overhead is $O(\sqrt{N})$ and replaced the previous protocol whose overhead is $O(N)$ by the new one. In addition to implementing searching module, we also implemented authentication module, secure communication module and GUI both for user and server applications to construct an entire searching system. We have installed the implemented system in the human genome databank to test an efficiency of the system.

2) Privacy-preserving genome sequence search

We designed and implemented a novel protocol which enables to search a similar genome sequence from a database. The method is based on a new approach that combines efficient string data structures such as the (positional) Burrows-Wheeler transform with a cryptographic technique called an oblivious transfer. The proposed method is order of magnitude faster than existing algorithms for finding substring match. In an experiment using 2184 aligned haploid genomes from the 1000 Genomes Project, our algorithm was able to perform typical queries within $\fallingdotseq$4.6 s and $\fallingdotseq$10.8 s for client and server side, respectively, on laptop computers. We also published the source code of the proposed algorithm at gitHub. The proposed method can be used in a wide range of applications.

3) Long-term preservation of the personal genome

The genome is inheritable to offspring and thus the protection period of genomic data could be very long. Despite that, most of the conventional studies use security parameters of the cryptosystem which were originally designed to handle other types of personal information such as a bank account and a phone number whose ideal protection period is much shorter than that of genome information. In our study, we address the problem of long term genomic data protection and suggest a novel approach that combines an information-theoretically secure method and a computationally secure method. We targeted on an allele frequency search problem which is modeled by 1 out of N oblivious transfer (1-N OT), and designed the novel 1-N OT protocol that achieves a long protection period while keeping utility.

In addition to above research results, we organized a workshop PRIVAGEN 2016 together with Finnish researchers. The workshop aimed to facilitate discussion among researchers in diverse fields including bioinformatics, genome ethics, machine learning, data-mining and cryptography. The PRIVAGEN was held as an official satellite workshop of GIW/InCoB 2016 which was the international conference of Bioinformatics, and there were 80 participants from more than 10 countries. We also organized a session for discussing the effect of new legislation of personal information in Japan on genome information analyses at IIBMP 2015 which was the largest domestic conference of Bioinformatics in Japan.

In this fiscal year, we published four referred journal papers and one paper for a domestic conference, and won three awards (IIBMP 2015 Excellent Research Award, IIBMP2015 Best presentation award, AIST president award (research)).