論 文 の 内 容 の 要 旨

論文題目

Probabilistic models for haplotype assembly and differentiation analysis

（確率モデルに基づくハプロタイプアセンブリ法と細胞分化機序解析法の開発）

氏　　名　　松本　拡高

The advancement of experimental technologies have enabled remarkable progress in molecular biology in recent years. However, the advancements in computational methods as well as sequencing technologies are essential to the progress of molecular biology. Molecular biological data frequently contains a mixture of multiple states and is hence heterogeneous, and computational methods are powerful tools to elucidate biological tasks from such heterogenous data. In this research, we accomplish the following two tasks, which cannot be investigated easily from experimental data, by developing computational methods. Firstly, we developed a computational method to infer individual haplotypes from sequencing data. Next, we developed a computational method to analyze single-cell expression dynamics during cellular differentiation.

1. Development of a probabilistic model for haplotype assembly

Haplotype information is useful for various genetic analyses, including genome-wide association studies. Determining haplotypes experimentally is difficult and there are several computational approaches that infer haplotypes from genomic data. Among such approaches, single individual haplotyping or haplotype assembly, which infers two haplotypes of an individual from aligned sequence fragments, has been attracting considerable attention. To avoid incorrect results in downstream analyses, it is important not only to assemble haplotypes as long as possible but also to provide means to extract highly reliable haplotype regions. Although there are several efficient algorithms for solving haplotype assembly, there are no efficient method that allow for extracting the regions assembled with high confidence. Therefore, we develop a probabilistic model, called MixSIH, for solving the haplotype assembly problem. Based on the optimized model, a quality

score is defined, which we call the `minimum connectivity' (MC) score, for each segment in the haplotype assembly. By using the MC scores, our algorithm can extract highly accurate haplotype segments. We also show evidence that an existing experimental dataset contains chimeric read fragments derived from different haplotypes, which significantly degrade the quality of assembled haplotypes. Therefore, we developed a method to detect chimeric fragments. The basis of our method is that a chimeric fragment would correspond to an artificial recombinant haplotype and would, therefore, differ from biological haplotypes in the population. We applied our method to two dilution-based sequencing datasets and the accuracy of assembled haplotypes increased significantly after removing chimeric fragment candidates.

2. Development of a probabilistic model for differentiation analysis

The advancement of single-cell technologies will shed light on the elucidation of the mechanism of differentiation. To fully analyze single-cell data, a novel computational method is necessary. There are several methods that use dimension reduction approach and reconstruct differentiation trajectory on the latent space to analyze single-cell expression data along differentiation. Although these approachese will be useful to extract the properties of differentiations, these methods have several problems such as the absence of standard in the selection of the axis. In this research, we developed a novel method SCOUP to analyze single-cell expression data along differentiation by representing the expression dynamics with Ornstein-Uhlenbeck process. In our evaluation, SCOUP can infer the degree of differentiation of a cell (pseudo-time) with high accuracy comparing to previous methods, especially for single-cell RNA-seq. We evaluated the cell lineage estimation and SCOUP can estimate more accurately than previous method, especially for cells at an early stage of bifurcation. To understand cell fate decision mechanisms, it is important to analyze cells immediately after bifurcation. We also developed a novel correlation calculation to analyze gene regulatory relationship while removing the spurious correlation. Thus, SCOUP will be a promising approach to analyze single-cell expression data during cellular differentiation and to elucidate regulatory mechanism of differentiation.