論 文 の 内 容 の 要 旨

論文題目　Inferring Chromosome Structures with Bidirected Graphs Constructed from Genomic Structural Variations
（ゲノム構造多型に基づく双方向グラフを用いた染色体構造推定）

氏　　名　安田知弘

The analysis of mutations in genomes are necessary to understand biological functions of genomes. Among various types of mutations, structure variations (SVs) are large scale mutations typically larger than 1 kb, and are attracting attentions. Examples of SVs include large deletions, insertions, inversions, translocations, and copy number variations. Next generation sequencing (NGS) technologies have made it possible to exhaustively detect SVs in genomes of thousands of individuals, including patients of cancers and serious congenital diseases. In addition to NGS technologies, a recent sequencing technology that can determine thousands of contiguous bases at once is now available. By using this technology, it would be possible to more easily and accurately obtain genome sequences in de novo manner.

To understand the impact of SVs on biological functions and infer possible mechanisms that caused SVs, it is vital to develop computational methods that accurately detect existence of SVs and their positions in the genome by using genome sequences obtained by these sequencing technologies. Because of the importance of SVs, a number of methods have been already proposed for detecting SVs by using NGS sequences. Although detected SVs should be further used to analyze their impact on genomes, computational tools that annotate detected SVs have not been as intensively developed as detection tools. In particular, although SVs are only local information in genomes, computational methods that utilize detected SVs to infer global structure of chromosomes are not well established. This can hamper our understanding of the effect of SVs, e.g. structures of proteins or regulations of genes affected by SVs. In addition, accurate detection of SVs are still difficult problem even with many existing methods, because the length of NGS sequences are limited to only a few hundred bases and thus detection of SVs involves finding complex patterns hidden in an enormous number of alignments of NGS sequences with the reference genome. Determining positions of SVs is obviously more difficult than determining existence of SVs. Nonetheless, knowing accurate positions of SVs clarifies the range of genomic sequences affected by SVs, and is also necessary for inferring the biological significance of SVs

or detecting mechanisms that had generated SVs. Most of existing methods take alignments of NGS sequences and the reference genome as input, and detect aberrant patterns of alignments as signatures that indicate existence and positions of SVs. Such signatures include aberrant distances and/or strands of alignments between paired reads obtained by NGS and the reference genome, aberrant number of aligned NGS sequences in a specific genomic region, and fragmented alignments. Because these signatures have already exploited, a new signature should be used to improve the existing methods. Because using NGS sequences to detect SVs include a computationally intensive step of mapping, i.e. aligning NGS sequences with the reference genome, accelerating this step is also crucial for exhaustive SV detection. In this thesis, we address these problems.

First, we address the problem of reconstructing global structures of chromosomes by using detected SVs. The problem had been previously formulated by Oesper et al. as an optimization problem on a graph constructed from SVs, which can be solved by calculating an optimal flow on the graph. However, they did not deeply analyzed computational complexity of the problem. In addition, the number and the length of chromosomes had not been considered. We formulate a new problem termed as the chromosome problem (ChrP) that takes into account the number and the length of chromosomes as well. In addition, we prove that ChrP is NP-complete by showing that there is an upper bound on the size of chromosomes in an optimal solution, and by reducing the Hamiltonian Cycle problem to ChrP. We also propose a biologically meaningful restriction on instances of the problem, termed as the weakly-connected constraint (WCC), and a variation of ChrP, termed as ChrW. ChrW imposes WCC on instances and removes limitation on the length of chromosomes. These modifications allow ChrW to be solvable in polynomial-time. Moreover, to show that removal of limitation of the length of chromosomes is necessary, another variation of ChrP that only imposes WCC on instances is defined and is proved to be NP-complete. Our result establishes a theoretical foundation of software tools emerging for the analysis of global structures of rearranged chromosomes. In computational experiments, our algorithm that solves ChrW was confirmed to be able to reduce noise in simulated SV data that include modified copy number variations (CNVs) or false positive translocations.

Second, we propose a method ChopSticks that accurately predicts positions of homozygous deletions, which is one of various types of SVs. ChopSticks mainly improves positions of homozygous deletions detected by finding alignments of paired reads with aberrant mapping distances. The paired reads with such alignments are called discordant reads. Positions of deletions calculated by using only discordant reads involve ambiguity. To reduce this ambiguity and to narrow down positions of boundaries of homozygous deletions, ChopSticks takes into account normally mapped sequences that have not been fully exploited by other methods, in addition to discordant reads. We theoretically prove that the expected distance between true positions and positions predicted by ChopSticks is close to distances of previous methods applied to NGS sequences with doubled depth of coverage. Experimental results also witnessed that our method is useful to predict accurate positions of homozygous deletions detected not only by using discordant reads but also by detecting drops of copy numbers.

Moreover, toward faster mapping of NGS sequences to the reference genome, we port widely used mapping programs to a many-core processor XeonPhi. In a computational experiment, the performances of the ported programs increased as the number of threads increased up to at least 60. This result indicates that concurrent execution of tens of threads on a many-core is promising for future performance improvement.

In this thesis, we also address the problem of detecting SVs by comparing multiple genome sequences constructed by de novo assembly. Such a method will be useful when the new technology that can generate long sequences becomes widely available in the future. Assuming that the input sequences are concatenations of a hidden set of sequences, our method infers the hidden sequences from the concatenations. To this end, we define a class of strings, called disjoint common substrings (DCS's). DCS's are similar to hidden strings and are nonetheless efficiently identified from given concatenations. Our algorithm identifies all DCS's in time linear to the total length of given concatenations. The effectiveness of our method were confirmed by a computational experiment.