論 文 の 内 容 の 要 旨

論文題目 Development of Alignment-free Algorithms for Various Post-genomic Data
（多様なポストゲノムデータのためのアラインメントフリーな
アルゴリズムの構築）

氏　　名　　小野寺　拓

DNA, RNA and proteins are basic biopolymers that are used universally among almost all biological systems. Inferring the functions of these molecules from their sequential or structural information is one of the most important problems in computational biology. As the so-called next generation sequencing (NGS) technologies and molecular dynamics simulation technologies are developed, it has become realistic to comprehensively investigate the space of sequences/structures that have never been observed. To advance such studies further, we need to develop efficient data analysis methods that are not dependent on alignment and more sophisticated (*de novo*) genome assemblers that can handle large and complex genomes such as metagenomes or eukaryotic genomes. We study these problems in this thesis.

In the study of alignment-free data analysis methods, we investigate the possibilities of alignment-free annotation methods for sequences and structures. Classification by composition-based string kernels and support vector machines (SVMs) is an existing alignment-free method for function prediction. Particularly, the spectrum kernel achieves relatively high classification accuracy and highly efficient computation based on the suffix tree. Thus, we start from this method and design better methods by applying more advanced data structures developed in string algorithms community.

In Chapter 3, we propose the *b*-suffix array data structure to make it possible to compute the kernel function that we introduce in Chapter 4 in time independent of the dimension of the feature space. This data structure is a generalization of the suffix array and it can also be applied as a string index that supports the search of patterns with wildcards in predetermined positions. Such pattern matching problems arise in spaced seed-based sequence homology search. We also propose non-trivial construction algorithms for the *b*-suffix array.

In Chapter 4, we propose the gapped spectrum kernel, a string kernel that is based on the frequency of substrings as well as the spectrum kernel. Different to the spectrum kernel, we introduce gaps (wildcards in pattern matching) to the substrings. A similar idea is used in an existing method called wildcard kernel. The wildcard kernel achieves high prediction accuracy by considering all gap patterns of a given pattern length and weights, but it was not known what happens when multiple but not all gap patterns are used. Applying the results from Chapter 3, we propose an efficient algorithm to calculate the gapped spectrum kernel and an algorithm to make a SVM prediction in time independent of the size of the support vectors. We also show that the sum of the gapped spectrum kernels corresponding to a given length and weights matches the wildcard kernel. From this relationship, efficient methods for wildcard kernel computation and prediction are derived. We also experimentally show that the sum of a few gapped spectrum kernels corresponding to randomly chosen parameters can predict protein families comparatively accurately as the wildcard kernel.

In Chapter 5, we study protein structure analysis. Protein structures are more directly related to functions than sequences are and thus, if available, they can be important clues to infer functions. On the other hand, structural alignment, the *de facto* standard method to measure structural similarities, is more computationally expensive than sequence alignment. Thus, in Chapter 5, we give an alignment-free kernel for protein structures applying the techniques for alignment-free kernels for sequences we saw in Chapter 4. Also, we propose an efficient method for kernel computation, which is based on an existing data structure called the two-dimensional suffix tree, and prediction method that takes time independent of the size of support vectors. We experimentally show that, compared to the most accurate similar existing method, the proposed method can achieve comparative accuracy while it runs more than 300 times faster.

We consider genome assembly problem in Chapter 6. Most existing genome assemblers first construct a graph that represents the overlaps of reads and try to recover the original sequence or long substring of it by following a path of the graph. While graph construction has been studied extensively, there is no established method for the part of recovering the original sequence from the graph. Particularly, one big open problem is how to process substructures introduced into the assembly graph by sequencing errors, repeat regions, diploidy/polyploidy or possibly other reasons. Existing methods detect such substructures by using simple motifs. Sometimes, however, complex substructures that cannot be detected by simple motifs considered in previous work do appear. What is required ultimately is to determine how to process these substructures but, different to

simple motifs, detecting such complex substructures is already non-trivial. We, therefore, give a graph theoretic characterization of these complex substructures, which we name superbubbles, and clarify several properties of them. We also propose an efficient algorithm to detect all superbubbles in a given graph. The algorithm takes time quadratic to the number of vertices in the worst-case, but it runs very efficiently in practice. We show the algorithm runs in linear time in expectation under a probabilistic model.

As a whole, in this thesis, we develop alignment-free algorithms to facilitate comprehensive studies of biological sequences and structures.