論文の内容の要旨

Arapan: A Systematic and Automated Genome Assembly System
(Arapan: 自動的ゲノム・アセンブリ・システム)

氏名　　モハッメド　サヒリ

Recent discoveries have been shown that DNA molecules became one of the most fundamental information in genetic sciences. Knowledge of DNA sequences will be ultimately fruitful for curing complicated diseases as the case of cancers for example. Decoding the code of life (i.e. DNA) into human natural languages will eventually lead us to new findings in various fields including medicine, biology, forensic biology, biochemistry, biotechnology and agriculture.

To obtain the genome sequence of any species needed what is now called sequencing technology. The complexity of DNA molecule structure as well as its tininess prevents any current sequencing technology to sequence the whole genome at once. As a result, we could not obtain the complete genome sequence but a set of reads that include sequencing errors and sometimes contaminated with cloning vectors or other foreign genomes. Bringing these reads all together in order to reconstruct the original genome sequence is called the (whole) genome assembly problem in literature. The different challenges that came along with DNA reads pushed computer scientists to design and develop more sophisticated assembly algorithms. This research topic epitomizes the backbone of this thesis. As a result, we developed an automated and systematic genome assembly system that was divided into two important sets of tools: preprocessing tools and assembly tools.

Because of the hugeness of sequencing data and the difficulty of other different constraints, we aimed at designing different tools for preparing the data to be independent from any genome assembler. Consequently, we could universalize the input of our assemblers to be “theoretically” adapted for any sequencing technology. By this way, the output data can be used for other purposes as well, and whenever a new sequencing technology appears in the market, we can just create simple tool for converting its data to the universalized input of our assemblers. More so, we can concentrate on developing the assembly algorithms regardless of the characteristics of the data produced by any sequencing technology. This separation can lead us to build more robust assemblers and creating new tools for preparing, analyzing, visualizing and validating the sequencing data. During this research, we could develop a sequencing errors corrector, trimming tool and other useful as well. In the first part of this thesis, we will explain in details the design and algorithms of each of these tools.

Most of genome assembly approaches use graph theory algorithms and data structures. Two approaches have been used to design the assemblers: the overlap graph-based approach and the de Bruijn graph-based approach. Each approach has some weak and good points. Creating more efficient assemblers requires analyzing the chosen graph more deeply so that it can be safely traversed for constructing long contiguous sequences. We discovered new substructures in the de Bruijn graph with the aim of developing a genome assembler that is able to produce more accurate results and avoid the misassemblies as much as possible. Assembly algorithms were kept simple, fast and designed in a cascaded way. As a result, we could develop two specialized de novo assemblers: Arapan-S and Arapan-M. We will explain step-by-step the overall architecture of our assemblers as well as the algorithms used in their different stages.

We will express and show the performance of each tool we created in the result section of each chapter throughout this thesis.