Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Informatyka Techniczna i Telekomunikacja

Rozprawa doktorska

Korekcja danych z sekwencjonowania genomów

Maciej Długosz

Promotor: prof. dr hab. inż. Sebastian Deorowicz

Gliwice, 2023

Correction of genome sequencing data

Abstract

Maciej Długosz
Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science
Department of Algorithmics and Software

Supervisor: prof. dr hab. inż. Sebastian Deorowicz

The dissertation focuses on a correction of reads obtained from a DNA sequencing performed with the machines of Illumina. To this day, many of algorithms were developed, aiming at a detection and elimination of errors present in a sequencing data. Algorithms efficacy and computational resources requirements vary significantly, causing a choice for a specified task to be difficult. Moreover, the existing overview papers do not give a comprehensive outlook, which would facilitate the decision.

Basing on an existing algorithms analysis, a new algorithm called RECKONER was developed. In the work, a set of requirements and principles of its operation were proposed. The algorithm was equipped with solutions allowing to improve a correction efficacy: an indel errors correction strategy, a method of a result choice from a possible solutions variety, a results verification strategy based on oligomers of two lengths. A new method of k-mer counting was introduced and a data structure being generated by KMC was utilized. Furthermore, an automated, basing on an empirical data method for a main algorithm parameter, k-mer length, determination was introduced. A computational complexity was determined. A number of the cases considered while the algorithm work was limited, reducing a computation time and a memory consumption. The algorithm allows for a parallel computation.

RECKONER was tested in experiments aiming at the efficacy evaluation and the resources consumption measurement. A comparative analysis with another algorithms was performed. Within that, sets of reads simulated with two methods were used, allowing to verify a number of eliminated errors and inspecting a potential of correction evaluation with such a data. In the experiments, many of read sets obtained from real sequencing processes were also utilized and an influence of the correction on *de novo* assembly and reads mapping tasks was observed. A similar analysis was performed with an original evaluation method, which is based on verifying an influence of the correction on a genome variant calling task. In the experiments, time, memory and scalability benchmarks were performed.

As a result of the experiments, it was concluded, that the possibility of the simulated reads utilization is limited. The evaluation method with a variant calling needs to be further examined. Finally, a group of the best correction algorithms was proposed to be utilized in the works related to a genome sequencing reads processing.