# Abstract

Modern Sequencing machines generate a large amount of data that must be processed. However, the existing algorithms and best tools are not perfect and produce archives . In this article, we will cover a compression algorithm called FQSqueezer that sequences data and processes single- and paired-end reads of variable lengths, and what was mentioned above was based on the dynamic Marco encoder algorithm.

The disadvantage of the proposed method are large memory and time requirements.


# Introduction

Genome sequencing has advanced to the point that it is now a mature technology with various medical applications. Illumina's instruments generate the vast majority of available data at a low cost low (e.g., about one thousand of U.S. dollars for whole human genome sequencing) compared to other sequencers. The sequenced reads are short (up to a few hundred bases) but with excellent quality, on the other side The PacBio or Oxford Nanopore 3rd generation instruments can produce much longer readings, but with worse quality and lower throughput. Unfortunately, items like short-distance repetitions of sections of data are rare in read sets, so researchers looked for other options. They initially concentrated on the compression of bases to collect reads from near regions of genomes.

Our algorithm, FQSqueezer, make use of the ideas from the prediction by partial matching (PPM)19,20 and dynamic Markov coder (DMC)21 general-purpose methods .

Some suggested techniques : the organization of the huge dictionaries, design of the PPM-like estimation of probabilities, use of a custom DMC as the stage following the PPM, the technique for prediction and correction of sequencing errors, technique of ordering the reads and making use of shared prefixes.

The main aspect of FQSqueezer is its compression ratio, usually much better than of the state-of-the-art competitors, i.e., FaStore12, Spring14, and Minicom15. however, our tool has  some drawbacks in terms of speed and memory usage" it is a few times slower than the mentioned competitors in compression and much slower in decompression.