Complexity Analysis of Algorithms: A case study about Bioinformatics Tools

Luan Ribeiro Siqueira, Victória Cardoso dos Santos, Mônica Silva de Oliveira, Adonney Allan de Oliveira Veras, Gislenne da Silva Moia

UNIVERSIDADE FEDERAL DO PARÁ

Abstract

The diverse analysis performed by omics sciences, driven by the reduction of costs in the DNA sequencing and reduction of the total time to carry out this process, resulted in an exponential increase in the deposit of all this information in public databases, for example, the National Center for Biotechnology Information - NCBI. The volume of data produced by these sequencing platforms demanded the development of algorithms capable of performing the most varied analysis, such as remotion of redundancy in raw reads from the sequencing process. However, it is worth mentioning that the existence of these various tools which performs this task with proven accuracy through their scientific publications, they do not analyze criteria related to the algorithmic complexity involved in their development. Therefore, this work demonstrated an analysis of algorithmic complexity, through empirical analysis already described in the literature, this analysis was performed with sixteen raw reads datasets with sizes ranging from 900 thousand to 12 million, they were obtained in the NCBI database in the Sequence Read Archive (SRA) format, they were converted to the FASTQ standard through the fastq-dump tool, the selected tools were: MarDRe, NGSReadsTreatment, ParDRe, FastUniq, and BioSeqZip. The analysis was performed on the R statistic platform, by using the GuessCompx package using the processing time of all datasets required by each tool as input, the models created were submitted to the glm adjustment function, in order to identify the function that indicates the complexity observed in each model. To this end, seven Big-O notations were observed:O(n), O(log(n)), O(n^2), O(n^3), O(1), O(n-log-n) and O(2^n). With the analysis of the results plotted graphically, it can be concluded that the NGSReadsTreatment tool obtained the least complexity in the processing of the datasets used in this analysis, presenting a linear complexity behavior, which leads us to infer that for datasets with high volume, this tool shows an interesting alternative to performing data processing.

Funding:
Link to Video: