In the eukaryotic genomes, elements such as protein-coding genes, telomeric sequences or centromeric sequences can be distinguished. The latter are necessary during cell division, because this is where the spindle fibres responsible for the proper division of genetic material are attached. Centromeric sequences often consist of thousands of tandemly repeated stretches of DNA, each very similar to the others. This property made it impossible to correctly read the centromeric repeat sequences in the past, and although the Human Genome Project had already published the "complete" human genome sequence in 2001, the centromeric sequences had to wait until 2022, when new technologies made it possible to read these complex regions. For this reason, centromeres still remain the "black hole" of the genome, and their study is severely limited by the lack of tools and experience. This is also a very important structure, because abnormalities in centromere sequences are associated with the development of some cancers and with abnormalities in the number of chromosomes, as in Down syndrome. Apart from its medical significance, the observed variability in the structure of centromeres is an important point in the study of many phenomena fundamental to biology, such as evolution of distinct species, and targeted modification of centromeres can be used in the production of new varieties in plant breeding.

Analysing centromeric sequences remains problematic, just as obtaining these sequences was. Classic methods of bioinformatic analyses often completely ignore repeated sequences because they were not designed for this purpose, and the repeated sequences themselves were not entirely reliable. Although sequences and basic analyses of single species are now available, there is still a lack of adequate classification and organization of the observed variability among eukaryotic organisms, and many potential mechanisms for the functioning and evolution of centromeres have been proposed on the basis of incomplete information about their sequences. There is therefore a need for centromere organisation classification and comparative analysis of centromere sequences and the related analyses of centromere proteins or epigenetic changes characteristic to centromeres.

In this project, bioinformatic tools enabling the analysis of centromeric sequences will be developed, taking as a starting point the previously published repeat sequence mapping software, TRASH, which will be used to perform large-scale analysis of centromeric sequences. The sequences for analysis will come from data shared by international consortiums involved in obtaining high-quality full genome sequences of many species.

In addition to providing important results on the function and organization of centromeres, as well as bioinformatics solutions for researchers describing genomic structures, this project may contribute to a better understanding of pathological phenomena occurring in the processes of carcinogenesis and chromosome segregation.