Silesian University of Technology
Faculty of Automatic Control, Electronics
and Computer Science

PhD Thesis

Methods for similarity analysis of low complexity regions in protein sequences

Patryk Jarnot

Supervisor: dr hab. inż. Aleksandra Gruca, prof. PŚ.
Co-supervisor: dr hab. Marcin Grynberg

Gliwice 2023

Abstract

In this doctoral dissertation, I focused on the problem of comparing Low Complexity Regions (LCRs) of protein sequences. These regions are important for many functions in proteins, such as RNA binding. However, scientists ignored them for a long time, focusing mainly on fragments of high complexity. This has led to the hypothesis that protein sequence similarity analysis methods are primarily designed for high complexity regions and are not efficient enough for LCR comparison. In the scope of this dissertation, I proved the hypothesis and proposed new methods for efficient LCR comparison.

The first step in creating new methods for analyzing similarity between LCRs was to check whether existing methods can handle them correctly. To test this, I searched for similarities between LCRs using BLAST, HHblits and CD-HIT, and analyzed obtained results. The selected methods are considered the gold standard in this field, and many other methods use the same or similar statistical models. The results revealed that some solutions used in these canonical methods are only efficient for comparing high complexity regions and are suboptimal for LCR comparison even after optimizing their parameters. I used the insights from the analysis to develop LCR-BLAST and GBSC methods. LCR-BLAST is a method derived from BLAST that has been adapted to search for LCRs, and was developed by adjusting BLAST parameters, simplifying scoring matrix, and adding a new metric for assessing similarity between sequences. GBSC, is a new method designed to identify and cluster clear and degenerate Short Tandem Repeats (STRs) in protein sequences. Both types of STRs cover a large part of the sequences considered as LCRs, leaving only a small part that lacks even blurred repeats. I also showed that this small fraction of LCRs missed by GBSC can be further cut down by reducing the alphabet of canonical amino acids. I compared GBSC with other methods for LCR identification and with selected methods for protein sequence clustering. To demonstrate the usefulness of these new methods, I used them in three different LCR analyzes in proteins. In the first study, I used GBSC to show that sequence assembly errors may occur in protein databases. In the next study, I used LCR-BLAST and GBSC to find K-D/E motifs that are similar to such motif found in RNA polymerase. Finally, both methods have also been used to find common LCRs in the SARS-CoV-2 and human proteomes to warn that selecting certain epitopes for drugs and vaccines can cause autoimmune disease.