RepeatPlus - program for finding repeats in nucleic acids and proteins

Ana Jelović[1], Nenad Mitić[2] and Miloš Beljanski[3]

[1]Faculty of Transport and Traffic Engineering, University of Belgrade, Vojvode Stepe 305, 11000 Belgrade, Serbia

[2]Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia

[3]Institute of General and Physical Chemistry, Bio-Lab, Studentski trg 12, 11000 Belgrade, Serbia

a.jelovic@sf.bg.ac.rs

RepeatPlus - program for finding repeats in nucleic acids and proteins

Introduction

Nucleic acids (NAs i.e. DNA and RNA) and proteins are linear biological polymers comprised of 4 or 20 "letters" or basic monomeric units. The monomeric units in NAs are nucleotides and in proteins they are amino acids. The lengths of NA and protein sequences can vary from short sequences of a couple hundred of base pairs or amino acids to hundreds of millions of base pairs or hundreds to thousands of amino acids. Often in these sequences shorter sequences that repeat themselves two or more times are found. These repeat sequences are called simply repeats.

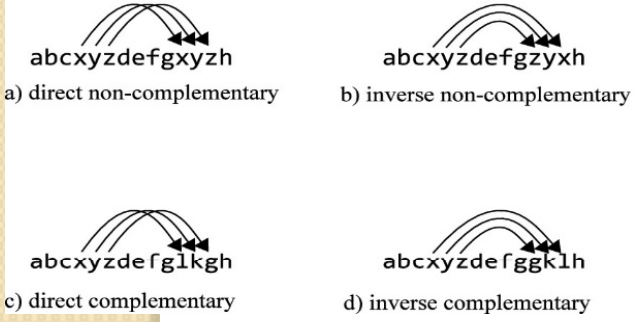

Figure 1. Example of repeat types

Repeat is defined as a pair of substrings in the input sequence that satisfies predefined conditions. Precise definition of repeats can be found in [1]. Depending on conditions, we have: direct and inverse repeats. Since NAs often consist of two complementary chains, for them two additional types of repeats can be considered: direct complementary and reverse complementary. Examples are shown in Figure 1.

Aim

As the number of repeats found can be very large, our goal was to provide a tool that can precisely find all repeats and filter them according to input arguments and outputs the results in a convenient way.

Method

To determine whether a repeat is statistically significant RepeatPlus uses the method described in [1] . With the given p-value and the expected number E of repeats the boundary above which the results are statistically significant is calculated.

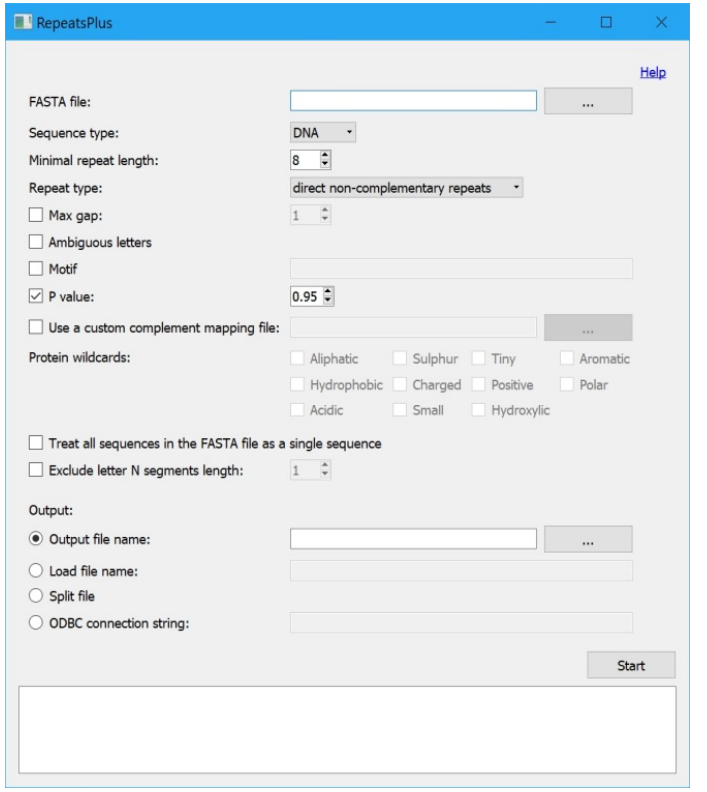

Figure 2. RepeatPlus window

Implementation

Program RepeatPlus consists of a Python graphical user interface and two batch processing console programs, StatRepeats and Repeats, that it executes based on the user input. RepeatPlus window is shown in Figure 2.

StatRepeats is a program that has three phases: finding all maximal repeats in an input sequence, calculating the expected number of their occurrences, and determining which of the found repeats are statistically significant in order to only output those.

Program Repeats finds all maximal repeats (in which the left and right side are not necessary identical due to the allowed number of mismatches) considering that some NAs or amino acids are treated as ambiguous letters. It uses an optimized brute-force approach. Repeats also finds maximal repeats that satisfy a predefined motif mask.

After finding all matching repeats RepeatsPlus can output results in many different ways. Console output and file output (in a number of different formats) are used most often, and are well structured so that they can be easily utilized by another program. Alternatively, the program can output repeats to an ODBC-compliant database, thus enabling simplified Data Mining or other Big Data approaches. There are also two other output options.

Options for treating multiple fasta files as one, option that provides mappings for different characteristics of amino acids (such as aromatic, charged, polar etc.) is also available. User can also define a custom complement mapping that is used over an arbitrary alphabet. Detailed explanation can be found in help and supplementary material.

Conclusion

Proposed filtering enables data reduction to set of statistically important (repeated) sequences which is extremely useful especially in the case of processing large number of repeated sequences with relatively small length, and provides a significant gain in performance and quality of results compared to outputting all the found sequences.

References

1. Jelovic A., Mitic N., Eshafah S., and Beljanski M.: Finding Statistically Significant Repeats in Nucleic Acids and Proteins, Journal of Comp. Biol. 2017, DOI: 10.1089/cmb.2017.0046