Patterns in scrambled chromosomes of ciliates
Denys Kukushkin, Department of Mathematics and Statistics

Abstract

Taking ciliates as model organisms, we study homologous DNA rearrangement processes. We use graphs and abstract words to capture patterns in thousands of scrambled genes of a recently sequenced genome. We observe common patterns that can explain complexities of all scrambled genes.

Methodology

- Precursor contains gene segments in a scrambled order. Product has gene segments arranged in the right order with non-coding DNA removed
- Mark repetitive sequences at the ends of each gene segment with letters
- Construct graph by connecting similar letters
- Traverse graph starting at this point
- Write down the abstract word

Recombination patterns in graphs and abstract words

- Given a precursor gene we build a corresponding graph and abstract word. For example, consider Contig20991.0.2\textsuperscript{1} of Oxytricha trifallax

Step 1
- Such DNA regions are removed first during rearrangement process\textsuperscript{4}. We remove graph loops

Step 2
- We identify longest sub-repeat word pattern and remove it

Step 3
- We identify and remove longest sub-word return pattern

Results and Conclusions

- The outlined process is used to estimate the complexity of the scrambled genes in the sequencing data of Oxytricha trifallax obtained by Chen, et al.\textsuperscript{2} by analyzing recombination graphs and abstract words of 15811 genes
- Step 1 showed that 13084 genes correspond to “all loops” graphs. Hence, these genes do not contain any types of scrambling. As a result, 1893 genes are left after Step 1 to analyze
- Step 2 showed that 464 genes correspond to the repeat word pattern. The histogram on the left depicts the distribution of repeat words compared to the word length
- After Step 2, we are left with 381 genes to analyze
- Step 3 showed that 111 genes correspond to the return word pattern. The histogram on the left depicts the distribution of return words
- After Steps 1, 2, and 3 only 215 remained to analyze for further complex recombination patterns

References