DOI 10.18699/SBB-2019-23

Context complexity of sites containing single nucleotide polymorphisms in human genome

Luzin A.N.[1]*, Dergilev A.I.[1], Tabikhanova Z.E.[2], Safronova N.S.[1], Orlov Y.L.[1, 3]
[1] *Novosibirsk State University, Novosibirsk, Russia*
[2] *Novosibirsk State Pedagogical University, Novosibirsk, Russia*
[3] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
** e-mail: a.luzin@g.nsu.ru*

Key words: bioinformatics, sequence complexity, medical informatics, polymorphism, SNP, databases

Motivation and Aim: Analysis of sequence text complexity is an approach for study genome structure based on sequencing data. We have analyzed sequence text complexity in flanking regions for the set of known single nucleotide polymorphisms (SNP) from the "1000 genomes" project. The aim was to find general and specific features for SNP sites associated with the diseases.
Methods and Algorithms: We used previously developed [1] and applied novel statistical computational methods to analyze genetic text based on its complexity. A complexity profiling in sliding window is applied to the sites, containing single nucleotide polymorphisms within a human genome. The complexity estimates were computed using previously developed program tool. This tool allows for both (i) complexity estimation of phased samples, and (ii) fast and effective defining the frequency spectrum of oligonucleotides with fixed lengths, and making a frequency comparison of oligonucleotides in different samples.
Results: A local decrease in text complexity level in SNP-containing sites is shown. Complexity profiles for SNP-containing sites shows that flanking monomer repeats define lower context complexity of sites containing SNPs within a human genome. An effect of local decrease in text complexity in SNP-containing sites is confirmed by analysis of polymorphisms in available model genomes (mice and rat).
Conclusion: Problem SNP sites analysis is of importance for personalized medicine and genomics studies. The changes in point mutation frequency were shown earlier for microsatellite containing sequences. Using extended data sets this work shows enrichment of polytracks and simple sequence repeats in local genome surroundings of SNP containing sites. We have found high frequent oligonucleotides within genome regions containing SNPs. Such oligonucleotides are related to nucleotide poly-tracks. The presence of poly-A tracks might be associated with an increased probability of double helix DNA breaks around mutable loci and following fixation of nucleotide changes.
Acknowledgements: Supported by the RFBR and ICG SB RAS budget project (0324-0019-0040).

References

1. Orlov Y.L., Te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. *J. Bioinform. Comput. Biol*. 2006;4:523-36.