DNA sequence complexity measures and analysis of short sequencing reads

Irina I. Abnizova

Welcome Trust Sanger Institute, Hinxton, Cambridge, UK
ia1@sanger.ac.uk

Rene te Boekhorst

University of Hertfordshire, Hatfield, UK

Arthur I. Dergilev, Fyodor M. Naumenko, Yuriy L. Orlov

Novosibirsk State University, Novosibirsk, Russia
arturd1993@yandex.ru; fedor.naumenko@gmail.com; orlov@bionet.nsc.ru

Key words: genome, text complexity, nucleotide polymorphisms, Hurst exponent, mapping

Here we discuss applications of text complexity measures to analysis of genomic DNA and problems of short-read mapping. The complexity of a symbolic sequence reflects the ability to represent this sequence in a compact form based on its structural features. A general approach to estimate complexity of symbolic sequences was suggested by A. N. Kolmogorov as early as in 1965 [1]. The Kolmogorov complexity is not a recursive function; it could not be realized in a computational scheme. However, for sequences of finite length, various constructive realizations of non-optimal coding were developed and applied for sequence analysis [1,2].

Competition of sequencing technologies, such as Illumina Solexa, SOLiD, PacBio leads to problem of data formats incompatibility, and more important, to non-reliable conclusion based on wrong DNA reads mapping to reference genome [3]. It was earlier shown that low complexity sequence regions, poly-tracks and simple repeats are related to systematic errors in genome sequence reads mapping and interpretation of sequencing results. The sequence complexity measures could be roughly dived to entropy estimates, linguistic complexity and algorithmic estimates including Lempel-Ziv compression method [2]. The measures of data quality are especially important for variant calling: in the particular case of SNP calling, a great number of false-positive SNPs may be obtained. We found earlier that not only the probability of sequencing errors (i.e. the quality value) is important to distinguish an FP-SNP but also the conditional probability of "correcting" this error (the "second best call" probability, conditional on that of the first call) [3]. Surprisingly, around 80% of mismatches can be "corrected" with this second call. We have developed several measures to distinguish between sequence errors and candidate SNPs, based on a base call's nucleotide context and its mismatch type.

Previously we used a complexity measure based on sequence segmentation, which we call complexity decomposition. A Hurst exponent, H, is considered as the measure of persistence of nucleotide occurrences in DNA stretch. In a case of random, identical and independent occurrences of nucleotides in DNA, H would be equal 0.5. For purine-pyrimidine alterations in DNA stretch, there long stretches of purine alternate with long stretches of pyrimidine, the sequence shows “persistence” and H will be larger than 0.5. Conversely, a low Hurst coefficient below 0.5 indicates “anti-persistence”, often caused by latent periodicity or short-range correlations (less than 10 bp). A high Hurst (>0.5) often indicates the presence of long-range (up to 10000 bp) correlations of DNA composition.

Text complexity estimates were formalized and coded in new software combining different methods. We show wide range of applications of such estimates for genome sequence analysis and nucleotide polymorphisms studies [4].

The work has been supported by RFBR (17-54-80103, 17-54-560028) and ICG SB RAS budget project 0324-2016-0003.

References

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

2. N.S. Safronova et al. 117 (2015) Analysis of SNP containing sites in human genome using text complexity estimates, *Journal of Biomolecular Structure and Dynamics*, **33**:sup1, 73-74.

3. I. Abnizova et al. (2012) Analysis of context-dependent errors for Illumina sequencing, *J Bioinform Comput Biol.*, **10**(2):1241005.

4. te Boekhorst R. et al. (2016) Computational problems of analysis of short next generation sequencing reads, *Vavilov journal of selection and breeding*, 20(6):746-755.