Applications of text complexity measures to genome sequences analysis

Nataly S. Safronova, Ekaterina V. Kulakova, Yuriy L. Orlov
Institute of Cytology and Genetics, Novosibirsk, Russia
orlov@bionet.nsc.ru

Low complexity regions are often treated as the regions of biased composition containing simple sequence repeats. New genome sequencing data issues the challenge to search for the regions with the low text complexity, which could be functionally important. Next generation sequencing technologies provides ever growing volume of sequence data to be verified, checked for technological errors and bias before functional annotation (Li et al., 2012).

The sequence enriched with imperfect direct and inverted repeats may be considered as the sequence with low complexity. Intuitively, complexity of symbolic sequence reflects an ability to represent a sequence in a compact form based on some structural features of this sequence. To evaluate text complexity, several groups of methods were developed: entropy measures (Shannon entropy), with the simplest of them using only the alphabet symbol frequencies; method of clusterization of cryptically simple sequences; evaluation of the alphabet capacity l-gram (combinatorial complexity and linguistic complexity); modifications of complexity measure by Lempel and Ziv; stochastic complexity; and grammatical complexity (Orlov&Potapov, 2004; Orlov et al., 2006).

We consider application of complexity measures to genomic texts. Intuitively, complexity of symbolic sequence reflects an ability to represent a sequence in a compact form based on some structural features of this sequence. The general approach to estimating complexity of symbolic sequences (binary texts) was suggested by A.N. Kolmogorov (Kolmogorov, 1965). Kolmogorov complexity is the length of the shortest code generating given sequence. Kolmogorov complexity is not recursive function (is not realized in computational scheme). However, for the sequence of finite length, various constructive realizations of non-optimal coding were developed (Lempel and Ziv, 1976), including classical compression algorithm LZ77. Approaches to genetic sequence analysis based on compression algorithms, has been suggested different authors. The mutual information, a measure intimately related to entropies, has been successfully used to predict protein coding regions in DNA (Grosse et al., 2000). We introduce a complexity measure based on sequence segmentation, which we call complexity decomposition, and we present several applications of this complexity measure. As the method for complexity evaluation, we have chosen the scheme of the text representation in terms of repeats, which uses the concept of complexity of a finite symbolic sequence, introduced by Lempel and Ziv (Lempel and Ziv, 1976), but oriented on DNA sequences. While studying complexity, we are interested not in a mere compression of genetic texts, but rather in detection of the regularities underlying it. The Lempel–Ziv complexity measure is based on text segmentation; so called complexity decomposition (Gusev et al., 1993). It may be interpreted as representation of a text in terms of repeats and allow find repeated sequence blocks, in smaller scale in gene transcription regulatory regions, and in larger scale in complete bacterial genomes. Original program is available at http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/ (Orlov&Potapov, 2004). Fast search of repeated fragments across mitochondrial genome using such complexity decomposition approach allows find "fragile" site potentially damaged by mutations (Guo et al., 2010). Recent application of complexity measures and analysis of over-represented oligonucleotides for transcription factor binding sites will be also presented (Putta et al., 2011).

As noted, when studying complexity of genomic sequences, we are not interested in a mere compression of these sequences, but rather in the detection of regularities hidden in these sequences, that could be in form of “building blocks” or overall similarity of sequence to other ancestor or “source” sequence. In addition, we apply the method of optimal text compression introduced by J.Rissanen in 1986 to construction of context source trees. It is a variant of Markov model of text generation allowing estimate probability for a given nucleotide sequence to be generated by fixed source sequence. The model allows find larger regions in genomes that differ by sequence content, but could not be detected by standard alignment-based sequence comparison method. We will discuss visual presentation of context trees for genomic sequences based on the method developed.

Finally we extent existing set of algorithms for sequence complexity estimations (Orlov et al., 2006) by new methods related to nucleotide variability in different positions, variance in nucleotide content composition to compare all mathematical measures of sequence complexity for different genomic applications related to transcription regulation (Putta et al., 2011). Our tool can investigate regions of lower complexity found in a phased sample of nucleotide sequences for presence of specific oligonucleotide signals and visualize it in complexity profile.

References

1) Grosse I., Herzel H., Buldyrev S.V. and Stanley H.E. (2000). Species independence of mutual information in coding and noncoding DNA. Phys. Rev. E 61, 5624-5629.

2) Guo X., Popadin K.Y., Markuzon N., Orlov Y.L., Kraytsberg Y., Krishnan K.J., Zsurka G., Turnbull D.M., Kunz W.S., Khrapko K. (2010) Repeats, longevity and the sources of mtDNA deletions: evidence from 'deletional spectra'. Trends Genet. 26(8):340-343.

3) Gusev V.D., Kulichkov V.A. and Chupakhina O.M. (1993) The Lempel-Ziv complexity and local structure analysis of genomes. Biosystems, 30(1-3), 183-200.

4) Kolmogorov A.N. (1965) Three approaches to definition of information quantity. Probl. Peredachi Inf. 1, 3–11. (in Russian).

5) Lempel A. and Ziv J. (1976) On the complexity of finite sequences. IEEE Trans. Inf. Theory, IT-22, 75–81.

6) Li G., Ruan X., Auerbach R.K., Sandhu K.S., Zheng M., Wang P., Poh H.M., Goh Y., Lim J., Zhang J., Sim H.S., Peh S.Q., Mulawadi F.H., Ong C.T., Orlov Y.L., Hong S., Zhang Z., Landt S., Raha D., Euskirchen G., Wei C.L., Ge W., Wang H., Davis C., Fisher-Aylor K.I., Mortazavi A., Gerstein M., Gingeras T., Wold B., Sun Y., Fullwood M.J., Cheung E., Liu E., Sung W.K., Snyder M., Ruan Y. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 148(1-2):84-98.

7) Orlov Y.L., Potapov V.N. (2004) Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res. 32(Web Server issue):W628-33.

8) 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.

9) Putta P., Orlov Y.L., Podkolodnyy N.L., Mitra C.K. (2011) Relatively conserved common short sequences in transcription factor binding sites and miRNA. Vavilov journal of genetics and breeding (Novosibirsk). V.15(4), 750-756.