StereoGene: a tool for fast correlation assessment and its application to the analysis of bivalent histone methylation

Elena Stavrovskaya, A.V. Favorov, Sarah Wheelan

Moscow State University, Leninskie gory 1-73, Moscow, 119992, Russia , stavrovskaya@gmail.com

Andrey A. Mironov

Institute for Information Transmission Problems , Bolshoy Karetny per. 19, Moscow, 127994, Russia mironov@bioinf.fbb.msu.ru

Vavilov Institute of Genral Genetics RAS , Gubkina str. 3, Moscow, 119333, Russia favorov@gmail.com

State Scientific Center Genetika , 1-st Dorozhniy pr., 1, Moscow, 117545, Russia

State Johns Hopkins University School of Medicine , 550 N Broadway ste 1103 Baltimore, MD 21205 USA

The modern high-throughput sequencing methods provide massive amounts of genome-focused, DNA-positioned data. This data is often represented as a function of the DNA coordinate (e.g. coverage). The genome- or chromosome-wide correlations between data from different sources may provide information about functional biological interrelation of the investigated features, e.g., transcription and histone modification. The key idea of the correlation studies is that two features that are similarly distributed along a chromosome may be functionally related. The correlation could also be treated as a function on genomic coordinate, and so we can not only assess the interrelations, but also to investigate their localisation inside the genome.

Previously, methods of correlation analysis were applied for numerical annotations and some biological results were obtained. But these methods do not allow to analyze positional correlations. The task to compute the spatial correlation was successfully solved only for interval annotations.

Here we present StereoGene that is a fast and powerful tool for estimation of correlations. Program implementation StereoGene allow to do analysis of two coverage profiles on human genome in 3-5 minutes. It works with quantitative and qualitative data. The program takes into account shifts of profiles relative to each other and search for correlation in somewhere around positions. It allows also to scale and sum profiles and compare profile combinations.

Besides the correlation and p-value for two input profiles StereoGene calculates the so called correlation profile, which is local correlation of input profiles for each genome position. The analysis of such a profiles reveals genome areas, where two features are highly (or on the contrary lowly) correlated. We applied this program feature to get bivalent promoter and bivalent enhancer regions in fetal and adult tissues. The former are identified as region with high local correlation of H3K4me3 and H3K27me3 histone methylation marks, the latter - H3K4me1 and H3K27me3. The analysis as being performed on pairs of fetal tissues and their mature ancestors allowed to get genes that change their activity during tissue development.
The work was supported by the Russian Foundation for Basic Research (№14-04-00576 and №14-04-01872).