Computer analysis of genome co-localization of transcription factor binding sites based on ChIP-seq data

Arthur I. Dergilev, Anatoly V. Svichkarev

Novosibirsk State University, 630090, Novosibirks, 2 Pirogova Str. arturd1993@yandex.ru

Yuriy L. Orlov

Institute of Cytology and Genetics, 630090, Novosibirsk, Lavrentyeva 10, orlov@bionet.nsc.ru

A scientific problem being solved is to study transcription factor binding sites (TFBS) co-localization in mammalian genomes using ChIP-seq data. Technology ChIP-seq, which combines chromatin immunoprecipitation (ChIP) and highly efficient DNA sequencing, allows to determine transcription factor binding sites in genome scale. The tasks of analyzing genome-wide ChIP-seq data rises are to identify the coordinates of TFBS and to compare their location with genomic annotation (relative location and distance to gene transcription start sites, promoter regions etc.). In addition to determining the location of binding sites for a transcription factor, there are problems of determining the cluster sites of different transcription factors, clusters together or located at a short (100-200 nt) distances on chromosomes assuming similar function and regulatory mechanisms. Programs processing huge amounts of text data (bed, wig files) identifying areas of intersection of genomic annotations (coordinates), adapted to the respective model genomes are technically necessary.

We developed set of programming script for TFBS location analysis. The study of clusters of sites ChIP-seq data on the status of binding sites of 15 different transcription factors in the mouse genome were used [1]. The computer program in C ++ language is developed to calculate the relative position of the coordinate TFBS and their clusters. Methods of establishing complex signals and patterns of the algorithm "Discovery" (program GeneDiscovery), previously developed in the framework of the theory of data analysis (Data Mining, Knowledge Discovery) in the context of signals DNA segments were used for the analysis of clusters of binding sites. We confirmed separation of TFBS clusters in mouse genome (embryonic stem cells) onto classes presented by Oct4, Nanog, Sox2 from one side, and c-Myc from another side. This analysis was extended to exact location of nucleotide motifs in ChIP-seq peaks relative to each other and iterative correction of such motifs.

The research has been supported by RFBR 14-04-01906 and ICG SB RAS budget project.

1. Chen X. et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells, *Cell*, **133**(6):1106-17.