Exploring protein occupancy patterns defining the specificity of interactions between DNA fragments

Artem Artemov

Department of Bioengineering and Bioinformatics, Moscow State University, 1-73 Leninskiye Gory, GSP-1, 119991, Moscow, Russia

Institute for Information Transmission Problems RAS

artemov@bioinf.fbb.msu.ru

Mikhail Gelfand

Institute for Information Transmission Problems RAS

Department of Bioengineering and Bioinformatics, Moscow State University

gelfand@iitp.ru

Alexander Favorov

VIGG RAS; GosNIIGenetika;

Johns Hopkins University School of Medicine

favorov@sensi.org

Andrey Mironov

Department of Bioengineering and Bioinformatics, Moscow State University

mironov@bioinf.fbb.msu.ru

Three-dimensional organization of chromatin plays a crucial role in genome functioning. Particularly, eukaryotic transcription is widely regulated through interactions of promoter and enhancer regions located far in terms of genomic distance but collocated in 3D. Moreover, actively transcribed genes are believed to be spatially concentrated into few so-called transcription factories. The mechanisms which determine the specificity of interactions between DNA fragments remain mostly unknown. Several DNA binding proteins, e.g. CTCF, are known to be involved in the formation of long-distance DNA loops. We explored genome-wide similarities in occupancy levels of certain DNA-associated proteins between spatially proximal DNA regions to study a potential role of these proteins in defining DNA-DNA interactions inferred from ChIA-PET experiment [1].

The results based on ChIP-seq data can at least partially be explained by the similarity of the methods which reveal protein occupancy and 3D packaging of DNA. Both of the methods start with DNA-protein cross-linking. This might cause a protein to be cross-linked not with

the DNA region of its original binding but with a region which happened to be spatially proximal. In order to exclude these potential effects, we first performed an analysis of similarities in occupancy based solely on transcription factors binding motifs [2] found in the DNA sequences.

We next compared the experimental occupancy levels from ChIP-seq experiments (ENCODE project, [3]) while DNAse hypersensitivity (ENCODE project, [4]) and sequence motifs data [2] was used to control false cross-linking. True sites of protein binding were assumed to be associated with DNAse hypersensitivity peaks which moreover were likely to harbor a motif for binding of this protein. We checked if presence of "true" binding site (i.e. a ChIP-peak for a particular protein within a DNAse hypersensitivity region harboring a binding motif) in one of two interacting DNA regions was associated with the presence of "true" site in its interacting partner compared to the presence of "false" sites (i.e. a ChIP-peak in a DNAse hypersensitivity region not harboring a binding motif or outside of hypersensitivity region).

To explore if tissue-specific contacts between DNA elements could be regulated by presence or absence of some DNA binding proteins, we used chromosome conformation and gene expression data for multiple cell lines. We selected the interactions between DNA regions which could potentially be put together by tissue specific transcription factors (e.g., estrogen receptor in MCF-7 cell line) and checked if presence of the corresponding TF in a tissue correlates with the proximity of these regions in 3D.

1. G. Li, X. Ruan, et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell,* **148**(1-2):84-98.

2. P. Shannon (2013). MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. R package version 1.2.2.

3. S. Neph et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints, *Nature*, **489**, 83–90.

4. R.E.Thurman et al. (2012) The accessible chromatin landscape of the human genome, *Nature*, **489**, 75–82.