Maximum Parsimony Interpretation of Chromatin Capture Experiments

Dirar Homouz, Gang Chen, Andrzej Kudlicki

Department of Biochemistry and Molecular Biology

Institute for Translational Sciences

University of Texas Medical Branch, Galveston, TX, USA

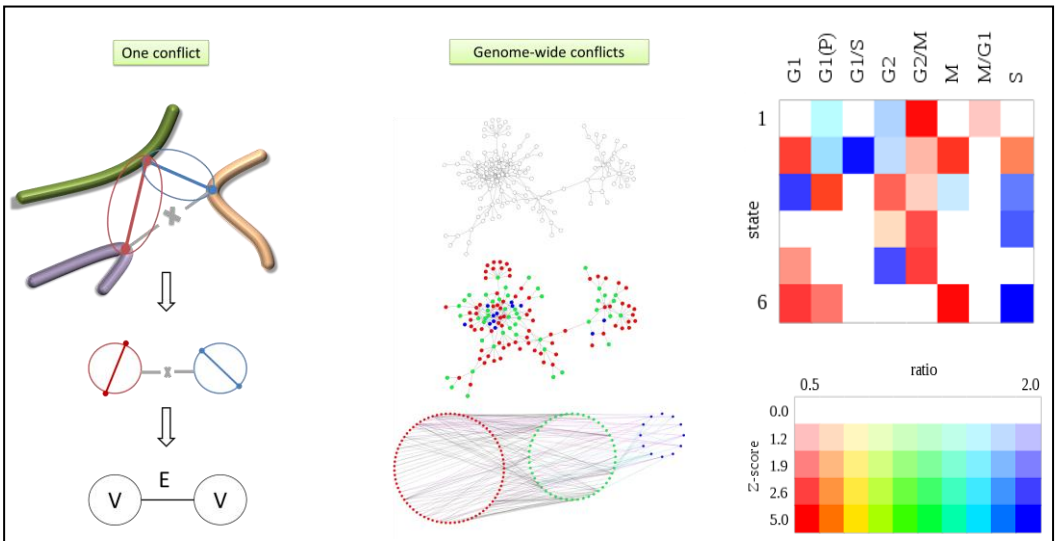
The three-dimensional conformation of the genome plays an important role in gene regulation at a wide range of scales, including the distribution of the nucleosomes, the folding of chromatin, as well as chromosomal conformation, chromosomal territories and interactions with the nuclear envelope. Chromatin conformation changes may be associated with all types of genomic and epigenomic variations, and have been implicated in a number of diseases, including cancer, neuromuscular and neuropsychiatric disorders, as well as other processes such as development and cell differentiation. Chromatin conformation capture techniques (3C, and its variants: 4C, 5C, Hi-C, 6C) probe the spatial structure of the genome by identifying physical contacts between genomic loci within the nuclear space; however, *existing methods of data processing provide no means of appreciating the variability between the cells in the sample.* Inhomogeneity of the experimental sample may reflect a number of different phenomena, as random thermal motions, subpopulations of cells executing different transcriptional programs, or different cell types present in the sample. Approaches have been developed that use stochastic simulations to create an ensemble of possible conformations, but these model-driven methods can only test certain global aspects of chromatin dynamics, and detailed insight into co-dependence of specific DNA contacts remains unavailable. Chromatin-capture data cannot be reliably interpreted in terms of a global conformation if DNA contacts from representing multiple chromatin states are present in the data.

We present a novel algorithmic framework that addresses this problem by analyzing the geometric and topological characteristics of an experimental DNA contact network. The approach is based on the observation that certain motifs in the 3C contact network can only be explained by inhomogeneity of the experimental sample. Specifically, in a haploid genome, a uniform conformation is not possible in which *locus A* is close to *locus B* and to *locus C*, but the Euclidean distance between the loci *B* and *C* is large: d(AC) >> d(AB)+d(BC), which leads to an impossible triangle. Such a situation corresponds to strong 3C signal for the AB and BC interaction, but no or few reads that would correspond to interaction between A and C. I the yeast genome data of Duan et al, we have identified up to $6.6*10^5$ such impossible triangles, involving $8.6*10^4$ DNA interactions, depending

on the conflict detection threshold. This result shows that cells with different genome conformations were present in the experiment, and that a large number of loci are affected.

To characterize the non-uniform sample, we employ the maximum parsimony principle, or Occham's Razor. We have developed and implemented a method of finding and characterizing the smallest set of homogeneous populations, which, mixed in the sample, will explain the experimental results. Our algorithm for optimal sample partitioning analyzes the topology of the DNA interaction network. Specifically a graph coloring procedure is applied to the graph representing conflicts between the observed DNA interactions, and interactions with the same color (or label) are interpreted as existing in the same state. Applying the method to the data of Duan et al, we demonstrate that the $6.6*10^5$ conflicts in the data can be resolved if the sample contains six homogeneous subpopulations of cells, each one with a different chromatin conformation state. GO enrichment analysis show that loci of genes from different functional categories form contacts in each of those states.

Our algorithms provide the first data-driven tool for studying the nature of the dynamic state of the interphase nucleus, its variability, and function. They will also allow building reliable models of the 3D nucleus and processing of data where different cell types were present in the experimental sample, and possibly lead to future diagnostic applications.



Partitioning the chromatin states of a haploid genome. Left: Representing conflicting interaction as edge in a line graph. Center: Coloring the graph representing conflicts yields groups of DNA interactions specific to each chromatin state. Right: DNA interactions in different states are significantly enriched in loci of genes active in different phases of cell cycle. inferred chromatin states (rows) affect distinct phases of cell cycle (columns). Color represents enrichment ratios (hue) and p-values (saturation).