A statistical approach to infer 3D chromatin structure

C.Caudai[1], E.Salerno[1], M.Zoppè[2] and A.Tonazzini[1]

1   National Research Council of Italy, Institute of Information Science and Technologies, Pisa
2   National Research Council of Italy, Institute of Clinical Physiology, Pisa

Our goal in the framework of the Italian Flagship Project *InterOmics* is to reconstruct a set of plausible chromatin configurations from Chromosome Conformation Capture data. To this end, we rely on a simulated annealing algorithm that samples the solution space defined by a data-fit function and a multiscale chromatin model. The data-fit only accounts for the largest, most reliable contact frequencies, in order to avoid deriving distances inconsistent with the Euclidean geometry. At each scale, the chromatin model consists in a chain of partially penetrable beads whose properties (bead sizes, elasticity, curvature, etc.) can be constrained through biochemical and biological knowledge. During the annealing process, the model configuration is evolved through quaternions rather than the usual Euler matrices, as this offers a number of advantages in terms of composition of successive perturbations and automatic satisfaction of the constraints. The output of the annealing scheme is not unique due to the degrees of freedom left by the geometrical constraints. This allows us to obtain multiple configurations compatible with both the data and the prior knowledge. We are validating our method by applying it to real Hi-C data from the long arm of the human Chromosome 1 [1]. The mean-square Euclidean distances computed from our results as functions of the genomic distances support previous experimental results indicating that highly expressed genomic regions are less compact than poorly transcribed regions [2].
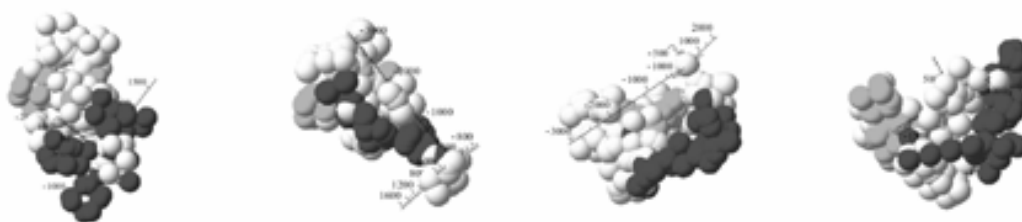

Figure 1. 4 outputs from Hi-C data from the long arm of chromosome 1 in human lymphoblastoid cells. Dark grey and light grey regions, both 3.6 Mb long, represent regions with high and low genic expression, respectively (as listed in the HTMseq website for a pool of similar cells [2]).

References

1.   E. Lieberman-Aiden, et al. Science. (2009), 326, 289.
2.   N. Versteeg, et al. Genome Research. (2003), 13:1998-2004.