E66: Claudia Caudai, Emanuele Salerno, Monica Zoppè and Anna Tonazzini. A multiscale model for 3D chromatin structure estimation using quaternions

Abstract: We present a method to reconstruct a set of plausible chromatin configurations from contact data obtained through Chromosome Conformation Capture techniques. We do not look for a unique configuration because the experimental data are not derived from a single cell, but from millions of cells. As opposed to most popular methods, we do not translate contact frequencies deterministically into distances, since this often produces structures that are not consistent with the Euclidean geometry. We build a data-fit function directly from the pairs of loci with the largest contact frequencies, assuming that they are likely to be in contact, and neglecting the pairs with very low or zero contact frequencies, as we cannot infer anything about their mutual distances. To obtain configurations consistent with both the data and the available biological knowledge, we introduce a chromatin model that can be suitably constrained. Taking advantage of the block structure of the contact matrix, we adopt a multiscale approach where the chromatin fiber is divided into a number of segments that can be treated in parallel. Each of them is modeled as a chain of partially penetrable beads whose properties (bead sizes, elasticity, curvature, etc.) can be constrained on the basis of biochemical and biological knowledge. The model parameters can easily be extended to exploit any further information available. Once the individual structures are reconstructed, each segment can be treated as an element of a new chain, and the procedure can be repeated recursively at different scales.

Our algorithm samples the solution space generated by the data-fit function through a Monte Carlo method. At each step, the subchains are perturbed by using quaternions. These are an extension of the complex algebra that offers a number of advantages, by avoiding singularities typical in the Euler matrix formalism, facilitating the composition of rotations, and allowing for a continuous evolution of the structure that is intrinsically compatible with the topological constraints.

To validate the new method, we applied it to real Hi-C data available online (Lieberman-Aiden et al., 2009). In particular, we analyzed the contact frequency data from the long arm of the human Chromosome 1 with a maximum resolution of 100 kb, obtaining a number of output configurations. For each configuration, the first division of the overall fiber included 25 topological domains (Dixon et al., 2012). The reconstructed structures were then assumed as single elements of a new chain (with nonuniform resolution), whose mutual interactions were estimated by the same algorithm. The output structures should be validated biologically. As a first test, we computed the relationships between the genomic and Euclidean distances of pairs of loci in the entire chains reconstructed. Our results are compatible with the analogous plots, derived from FISH experiments on the same genomic region, found in Mateos-Langerak et al. (2009).