Three-dimensional chromosome structures from energy landscape

Gamze Gürsoy[a] and Jie Liang[a,1]

The human genome contains about 2-m length of DNA and is packed into a small cell nucleus of approximately cubic-micrometer size. A central question in genome biology is to understand how chromatins are organized in such a compact volume, while biological functions such as gene expression, DNA replication, and DNA repair are robustly orchestrated. Over the past two decades, experimental studies based on chromatin fragmentation and proximity cross-linking have given us quantitative information on the frequencies of long-range interactions among genomic elements (1). With recent development of the Hi-C methodology (2), frequencies of such interactions are now known at 1-kB resolution (3). These studies lead to discoveries of finer organizational structures of compartments, subcompartments, and topologically associated domains (TADs) (3, 4). Although these structures have been inferred from analysis of heat maps of frequencies of genomic interactions (3, 5, 6), a grand challenge in studying the 3D genome is to gain mechanistic understanding of the general principles governing chromatin folding and their spatial organization. In PNAS, Di Pierro et al. (7) introduce an energy landscape theory and a predictive model of chromosome architecture.

Di Pierro et al. start by considering the roles of specific biochemical interactions. Although generic polymer models of chromatin have generated important insight into the overall behavior of chromatin, growing evidence suggests that biochemical interactions are critical for 3D genome organization (8). Di Pierro et al. assume that chromosomes fold under the influence of a cloud of proteins, which bind to different sections of chromatin with different affinities and specificities. To recapture the energy landscape governed by these interactions, Di Pierro et al. develop the minimal chromatin model.

The first ingredient of Di Pierro et al.'s model is the partitioning of the genome into intervals of a handful of types. Each interval type is characterized by its histone modifications and a characteristic combination of nuclear proteins it interacts with. As

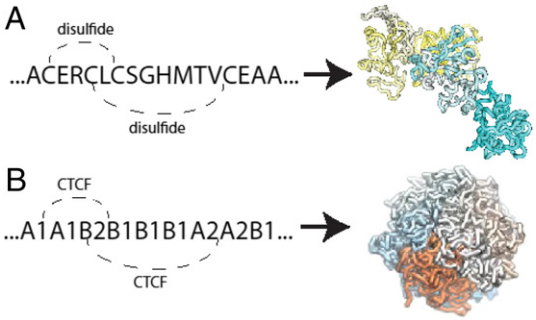

Fig. 1. Energy landscapes of protein and chromosome folding. (*A*) A protein sequence of amino acid residues and disulfide bonds between cysteine residues dictates the energy landscape of protein folding, which describes interactions between residues and solvent. The structure of the protein can be predicted from this energy landscape. (*B*) A sequence of genomic interval types and looping interactions between anchor loci of CTCF motifs dictates the energy landscape of chromosome folding, which describes interactions between genomic intervals and nuclear proteins. The structural ensemble of chromosome can be predicted from this energy landscape.

demonstrated by a number of biochemical and structural studies on *Drosophila* and human genome, distinct chromatin subcompartments corresponding to different interval types can be clearly identified, with exhibitions of characteristic histone marks and patterns of long-range interactions (3, 9). Di Pierro et al. model the effects of binding between two chromatin segments by approximating the free energy changes as a value that depends only on the types of the two contacting intervals. A similar approach has been successfully applied to study the formation of TADs (10). The second ingredient of Di Pierro et al. is that loops have high propensities to form between pairs of anchor loci, and the effects of loop formation can be modeled by changes in effective free energy incurred at the pair of loci where loops form. Di Pierro et al. assume that the anchor loci are mostly associated with CCCTC factors (CTCF) binding motifs. The important role of CTCF in loop formation has been

[a]Richard and Loan Hill Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607
Author contributions: G.G. and J.L. wrote the paper.
The authors declare no conflict of interest.
See companion article on page 12168.
[1]To whom correspondence should be addressed. Email: jliang@uic.edu.

well documented (8), and this assumption is also supported by experimental evidence that loops anchored at CTCF sites are readily recognizable from frequency heat maps of Hi-C studies (3). The third ingredient is the consideration of generic gain or loss of effective free energy when a pair of loci come into contact. This free energy change is independent of interval types and loop formations, and is only determined by genomic distance between the contacting loci. It models changes in local structures of an ideal chromatin. Di Pierro et al. summarize these considerations into an information-theoretic energy function, with one term and an associated coefficient representing each factor, along with an additional term representing properties of a generic homopolymer such as excluded volume effects. With this model formulation, Di Pierro et al. follow the principle of maximum entropy and generate ensembles of structures of a 2,712-bead model of chromosome 10 and iteratively adjust the model parameters. The ensemble generated from the final set of parameters can simulate Hi-C frequencies of the 136-Mbp chromosome (3) at 50-kb resolution, exhibiting an excellent agreement with Hi-C measurements (Pearson's $r = 0.95$).

A previous study of populations of chromatin structures based on Hi-C data also succeeded in reproducing interaction frequencies with similar correlation for the whole genome, albeit at lower resolution (11). The energy landscape of the minimum chromatin model differs as it is formulated to capture general principles at play in chromatin folding. Indeed, Di Pierro et al. generated an ensemble of chromatins of all other chromosomes using the minimum chromatin model, with parameters trained only on chromosome 10. With just 27 parameters, the heat maps of Hi-C–measured frequencies (3) of the rest of chromosomes unseen during parameter training can all be simulated with excellent agreement ($r = 0.95$) using annotations of chromatin types and CTCF looping locations from Rao et al. (3) as input.

The chromatin ensembles predicted with the minimum chromatin model of Di Pierro et al. reveal a number of interesting observations that have been reported in previous Hi-C and 3D-FISH studies (3, 12). Each chromosome is found to form a compact chromosome territory, with phase separations among different chromatin types observed within the territory. Similar phase separation was also observed when two interacting chromosomes were simulated simultaneously, suggesting that the formation of chromosome territories can be inferred from the same energy landscape that was used to simulate single chromosomes independently. In addition, subvolumes comprising a single type of chromatin interval are frequently found within a chromosome territory. Furthermore, highly expressed genes are often colocalized and predominantly lie in the less densely packed periphery of the chromosome territory. Di Pierro et al. also demonstrate the importance of biochemical interactions encoded in the model, as none of these observations appears in a control study when simple homopolymer chains with all biochemical information removed are used.

Di Pierro et al. further examine nonlocal properties of chromosome structures that cannot be directly inferred from Hi-C frequency data. Confined long polymer chains in equilibrium form knots with high probability. However, knotted chromatin structures would occlude access of *cis*-regulatory elements and transcription factors to genes, potentially hindering cellular functions. When Di Pierro et al. calculated knot invariants, a mathematical tool for knot detection, of sampled ensemble of chromatin conformations, modeled chromatin chains are found largely devoid of knots. Unknotted chromatin was a property that was used to justify the nonequilibrium fractal globule model (2, 13). Results of Di Pierro et al. show that an equilibrium mechanistic model based on minimalistic assumptions can generate ensembles of unknotted chromatin chains. Along with the presence of topoisomerases, enzymes that cut the double-stranded DNA to untangle knots, the formation of knots is unlikely to occur with significant frequency in cell. Furthermore, the power-law scaling relationship between the probability of contact formation and genomic distances, another justification of the nonequilibrium fractal globule model (2, 13), is also reproduced with great accuracy from the simulated equilibrium ensembles.

The energy landscape specified by the minimum chromatin model is reminiscent of the energy landscape theory developed in protein folding studies (Fig. 1). In place of a sequence of amino acid residues, we have a sequence of chromatin beads of different types. Loop formation between CTCF sites is analogous to the formation of disulfide bonds between cysteine residues. Admittedly, folded chromatins are quite different from folded proteins: contact interactions are likely to be transient, subpopulations of cells may have different chromatin conformations, and chromosomes are likely only partially structured. Nevertheless, the energy landscape theory for chromatin architecture may provide a general framework for answering important questions on 3D genome organization. As pointed out by Di Pierro et al., it is conceivable that, with a well-constructed energy landscape, 3D ensemble structures of whole genomes can be predicted computationally using 1D genomic data of epigenetic modifications, which can be cost-effectively assayed using ChIP-Seq, RNA-Seq, DNA-seq, and other epigenetic profiling techniques. The energy landscape theory of chromatin folding also helps to raise new questions. For example, one can ask how the "sequence" of genomic intervals change their labels or "mutate" during development and disease, how these mutations would lead to different 3D structural ensembles of chromatins, and how these changes would affect cellular phenotypes.

Energy landscape theory and the minimal chromatin model of chromatin folding by Di Pierro et al. provide a general framework for developing transferable, predictive, and physical models that can help to understand the mechanism of 3D genome organization.

Looking ahead, there are a number of technical issues that need to be resolved. It is still challenging to distinguish artifacts, generic polymer effects, and biologically specific effects in interaction frequencies measured in Hi-C data, so energy landscape of chromosome folding can be further improved and refined. The practice of aggregating Hi-C interactions into bins helps to identify domains and other finer structures from frequency maps but may introduce unwanted artifacts (14). Much of the measured genome-wide chromatin interactions in budding yeast (15) is due to generic polymer effects under the constraints of a few landmarks (nucleolus, spindle pole body, and centromere attachments) (16, 17), and identifying biologically important interactions is a nontrivial task (18). Furthermore, the abundance of CTCF motifs complicates the determination of loci parings for loop formations from 1D epigenomic data. In addition, inferring structural units of gene regulatory machineries that span just a few

kilobases requires chromatin models of finer resolution. These roadblocks, however, will likely be removed as the advancement in theory, model, and experimental measurements marches on. Energy landscape theory and the minimal chromatin model of chromatin folding by Di Pierro et al. provide a general framework for developing transferable, predictive, and physical models that can help to understand the mechanism of 3D genome organization. It is envisioned that computational models of 3D chromatin structures will help to decipher complex mechanisms involving higher-order genomic interactions that control cellular phenotypes. Successes in constructing predictive energy landscapes hold the promise of enabling biological discoveries such as identifications of novel enhancer–gene interactions through folding of 1D genome and epigenome into an ensemble of 3D chromatins.

Acknowledgments

Authors' research is supported by National Institutes of Health Grant GM079804, National Science Foundation Grant MCB1415589, and the Chicago Biomedical Consortium with support from the Searle Funds at The Chicago Community Trust.

1 Dekker J, Rippe K, Dekker M, Kleckner B (2002) Capturing chromosome conformation. *Science* 295(5558):1306–1311.
2 Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950): 289–293.
3 Rao SS, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–1680.
4 Nora EP, et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485(7398):381–385.
5 Fortin JP, Hansen KD (2015) Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 16:180.
6 Dixon JR, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380.
7 Di Pierro M, Zhang B, Lieberman Aiden E, Wolynes PG, Onuchic JN (2016) Transferable model for chromosome architecture. *Proc Natl Acad Sci USA* 113:12168–12173.
8 Phillips JE, Corces VG (2009) CTCF: Master weaver of the genome. *Cell* 137(7):1194–1211.
9 Filion GJ, et al. (2010) Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* 143(2):212–224.
10 Jost D, Carrivain P, Cavalli G, Vaillant C (2014) Modeling epigenome folding: Formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res* 42(15):9553–9561.
11 Tjong H, et al. (2016) Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci USA* 113(12): E1663–E1672.
12 Boettiger AN, et al. (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* 529(7586):418–422.
13 Mirny LA (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res* 19(1):37–51.
14 Ay F, Bailey TL, Noble WS (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res* 24(6):999–1011.
15 Duan Z, et al. (2010) A three-dimensional model of the yeast genome. *Nature* 465(7296):363–367.
16 Tjong H, Gong K, Chen L, Alber F (2012) Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res* 22(7):1295–1305.
17 Wong H, et al. (2012) A predictive computational model of the dynamic 3D interphase yeast nucleus. *Curr Biol* 22(20):1881–1890.
18 Ay F, et al. (2014) Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res* 24(6):974–988.