Free energy based high-resolution modeling of CTCF-mediated chromatin loops for human genome

Wayne Dawson[1] and Dariusz Plewczynski[1,2,3,*]

[1]Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Banacha 2c, Warsaw 02–089, Poland,

[2]Faculty of Pharmacy, Medical University of Warsaw, Banacha 1, 00-001 Warsaw, Poland

[3]Centre for Innovative Research, Medical University of Bialystok, Białystok, Poland

[*] corresponding author

Wayne Dawson w.dawson@cent.uw.edu.pl
Dariusz Plewczynski d.plewczynski@cent.uw.edu.pl

Abstract:

A thermodynamic method for computing the stability and dynamics of chromatin loops is proposed. The CTCF-mediated interactions as observed in ChIA-PET experiments for human B-lymphoblastoid cells are evaluated in terms of a polymer model for chain folding physical properties and the experimentally observed frequency of contacts within the chromatin regions. To estimate the optimal free energy and a Boltzmann distribution of suboptimal structures, the approach uses dynamic programming with methods to handle degeneracy and heuristics to compute parallel and antiparallel chain stems and pseudoknots. Moreover, multiple loops mediated by CTCF proteins connected together and forming multimeric islands are simulated using the same model. Based on the thermodynamic properties of those topological three-dimensional structures, we predict the correlation between the relative activity of chromatin loop and the Boltzmann probability, or the minimum free energy, depending also on its genomic length. Segments of chromatin where the structures show a more stable minimum free energy (for a given genomic distance) tend to be inactive, whereas structures that have lower stability in the minimum free energy (with the same genomic distance) tend to be active.

Introduction

In eukaryotic cells, detailed experimental identification of the structure of chromatin fiber inside of the cell has revealed considerable higher order organization and packaging in a hierarchical fashion [1-12]. Structural determination strategies have forged ahead with a number of high-throughput methods to obtain genome-wide maps of chromatin organization; e.g., chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and high-throughput chromosome conformation capture (Hi-C) [6,13]. Recent work has centered on including chromatin immunoprecipitation (ChIP) with ChIA-PET for specific protein factors (e.g., CTCF and RNA polymerase II) [2], offering structural information at a resolution of 1000 base pairs (1 kbp), given sufficient sequencing statistics. In general, a resolution of 1 kbp is now achievable as specific target proteins can be isolated – pushing whole genome structural resolution into the range of 50 nm. However, the experimental data reflects an ensemble of structures; i.e. the interaction data is collected typically from around 100 million cells, each having a different three-dimensional structure of nucleus. Moreover, chromatin itself is a dynamic system that can take on a variety of structures and thermodynamic states over time. Unlike many protein and RNA structures that often take on a rather definite average shape (or shapes), the structural features of chromatin are more plastic and require clustering to extract meta-structures. Therefore, identifying structural motifs of the chromatin chain and estimating the likelihood of particular motifs within the ensemble is now possible to explore.

Chromatin is a complex heteropolymer comprised of many different components. Double-stranded DNA (dsDNA) in eukaryotic organisms is bundled into packages called nucleosomes: a structure consisting of two full turns of the dsDNA (about 150 bps) around an octamer of histone proteins forming a core of the nucleosome. The octamer contains a tetramer core $(H3–H4)_2$ with a dimer of H2A–H2B capping each side of this wrap [14]. Connecting each octamer is a H1 subunit consisting of various subtypes. The globules of dsDNA/histones (the nucleosomes) form the chromatin fiber that comprises the genome. At the scale of *nucleosomes*, the structure of the chromatin fiber is thought to be somewhat random [9,15-18]; however, there also appears to be significant global organization [19-22]. The three billion base pairs (3 Gbp) of DNA of the human genome packaged as chromatin fiber could stretch out to roughly two meters [2,18] with a diameter of about 10 nm (e.g., PDB id: 5DNM, 5B31, 5KGF, etc.).

Here we analyze data at the resolution scale of 5 kbp beads, a scale where the beads comprising a chain can be treated within beads on a string polymer model; each 5 kbp genomic segment is represented as single monomer. The 3D topology of chromatin is thought to influence the regulation of gene expression and regions of active and repressed

transcription in the cell, where diverse parts of the chromatin fiber can be found in proximity of each other [2-4,23-26] in what are known as chromatin loops. One part of such a proximal chromatin segment contains the promoter(s) of a given gene, whereas the second part is enriched with enhancer DNA sequence(s), which amplify the transcription rate of a specified gene [27]. The interaction between the two ends of a loop is mediated by CTCF proteins that stabilize the interaction by forming a dimer with parallel orientation of both protein components [1,2,19]. Multiple loops can co-localize within the same three dimensional genomic loci; the process is mediated by many CTCF dimers that form together CTCF islands, or rafts (see Fig 1E). Such proximate collections of loops are defined here as chromatin contact domains (CCDs), similar to the term topologically associating domains (TADs), which is used widely in the field of 3D genomics. Such CCDs sizes often range between several kbp to several Mbp and show considerable similarity between cells and/or stages of cell development [2,3,23,24,26].

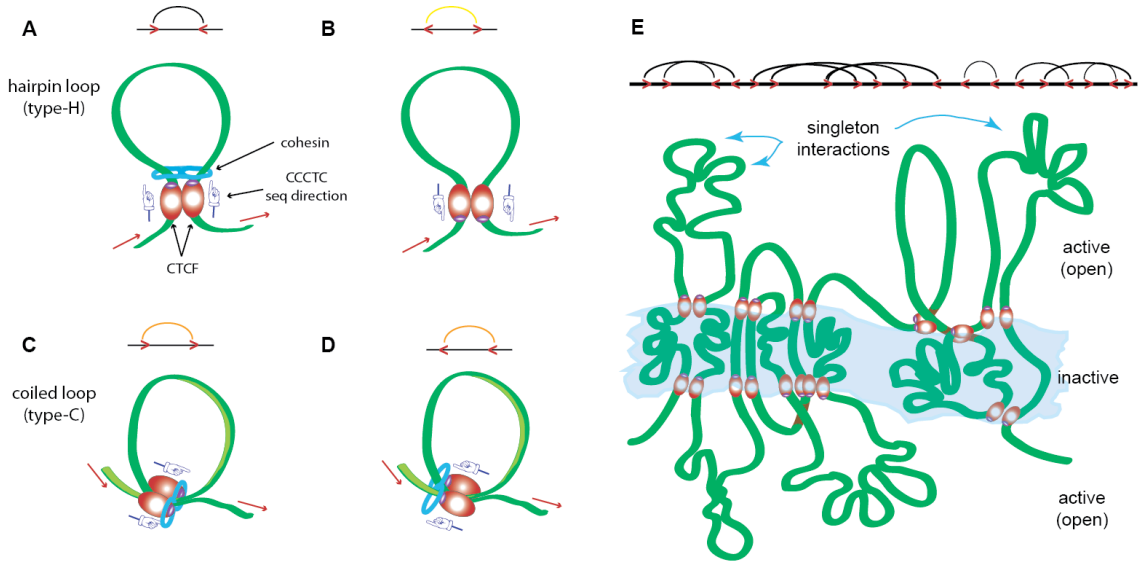

Figure 1. Examples of various types of loops formed by CTCF dimerization and its interactions with chromatin. (A) A convergent loop, (B) a divergent loop, (C) a tandem right loop, (D) a tandem left loop, and (E) combinations of A-D in the form of CTCF islands (a cartoon, but characteristic of regions like Chromosome 10). The 5′ to 3′ direction of the sense strand of the DNA sequence is indicated by the red arrows, cohesin is indicated by the blue feature enclosing structure, and the implied sequence direction for the CTCF motif is indicated by the directional pointers. The additional structural features are elaborated on in Figs S2-S4 (in Supplement).

The CTCF dimers combined with two cohesin ring-like multi-domain structural proteins binding in a parallel direction as a multi-protein complex. The attachment to the chromatin chain can occur in an antiparallel or parallel direction at the loop anchor points – depending on the direction of the zinc fingers binding in the underlying DNA sequence motif. The most frequent and strongest interaction is the convergent loop (Fig 1A), where the chromatin chain is anchored in an antiparallel direction with the CTCF dimer bound in a parallel direction so as to point into the loop (Fig 1A). Divergent loops [26] occur when the orientation of the CTCF dimer is in the opposite direction (Fig 1B). When the direction of the chromatin chain binds in a parallel interaction with the CTCF dimer, two types of structures of equal tendency and intermediate strength are suggested: tandem right (Fig 1C) and tandem left (Fig 1D) [2]. In general, 80% of the loops appear to be convergent [2]. These CTCF structure also appear to group into islands as shown in Fig 1E, where the central region is the largely insoluble and inactive part (heterochromatin) whereas the regions jutting out are more accessible to transcription factors and therefore active (euchromatin).

Therefore, we have opted to call these features structural motifs and formalize a notation for them; Fig 1 and Figs S2-4 (Supplement). Fig 1 consists of various types of CTCF-mediated loops and their combination into islands. Figs S2 through S4 (Supplement) show various types of singleton structural motifs or combination of singleton and CTCF or RNA Pol II interactions: simple loops, internal loops and multiloops as well as parallel and antiparallel stems and pseudoknots. The color scheme in Fig S2D indicates the strength of the interactions (black is maximum counts/minimum free energy).

In recent years, there has been considerable interest recently in finding ways to model the distribution of chromatin structure physically using various approaches such as population-based analysis [28] and polymer based models using molecular dynamics (MD) simulations [1,2,11,12,29,30] or Monte Carlo (MC) simulation techniques [8]. These models provide 3D structures of the chromatin. However, MD or MC simulation techniques do not guarantee an exhaustive search of such a large landscape for chromatin structural motifs, nor is it easy to say that a structure found in this way is actually the most stable structure [28].

In this work, based upon the observed contacts (referred to as the pair interaction frequency (PIF)) of the chromatin fiber contacts obtained from ChIA-PET data, we have developed a free energy based model with suboptimal structures that extracts the major part of the Boltzmann distribution of the ensemble of chromatin structural motifs distinguished in terms of singletons, CTCF (or RNA Pol II) binding sites and multimeric CTCF islands. The approach handles complex motifs such as pseudoknots (involving both parallel and

antiparallel chain interactions) using heuristic adaptations to the DPA. Based on the observed PIF ensemble found in heatmaps obtained from ChIA-PET [2], or *in situ* Hi-C experimental data [6,26], we transformed the PIF information into binding enthalpy and combined this with a highly flexible contact (cross-link) based entropy model that was shown to be rather successful for solving RNA secondary structure and pseudoknot structures [31-34]. We make here the underlying assumption that chromatin at the highest resolution (around 1 kbp resolution) can be characterized by a 2D representation of complex structural motifs as a kind of meta-structure not so unlike what is observed of proteins and RNAs.

Using CCDs, we were able to identify loops within the chromatin structures that tended to show activities consistent with such a free energy model. Hence, ensembles that have a dominant structure (a stronger free energy with large differences in the thermodynamic probability between the principal structure and neighboring structures in the list) tend to be inactive on the one extreme. Ensembles with several conformations of nearly equal weight –a weaker free energy (FE) and small differences in the thermodynamic probability between very diverse structures – tend to be active structures.

Results and Discussion

There are two objectives in this work. The first is to use polymer physics to obtain information on the dominant stable structural motifs of a given ensemble of observed contacts from a particular cell line (GM12878) and to discern the dynamics of the chromatin. The second is to use this information to identify regions of active euchromatin and inactive heterochromatin.

The contact map data obtained from experiments provides us with a picture of the collection of observed contacts (ensemble) within the large population of cells. The thermodynamics of polymers provides us with an understanding of the influence of any given collection of contacts on the observed structure (or structures) in the ensemble in the form of a statistical weight; i.e., the Boltzmann distribution. The resulting structures are 2D because what is observed is not one specific 3D structure, but individual contacts between diverse parts of the chromatin fiber. Using thermodynamics distinguishes the relative contribution from different contacts and the likelihood of particular structures. This in turn

permits some picture of the like dynamics of chromatin from the experimental data using the thermodynamic probability of different structures.

In this model, we assume that the observed frequency of contacts is identical for every cell. Most likely, when millions of cells are measured, each particular cell will have common housekeeping genes where the expression is identical, but particular states of the cell that reflect that particular conditions of that given cell. The configuration of these particular states is largely unknown. However, since the CTCF contacts are the source of the major interactions, in any particular loop study, it is possible to turn on and off these interactions and fit the observed ensemble of cells. This will be a matter that will be addressed in future work.

When one particular structure dominates all the other structures by a substantial margin, this means that most of the time, the structure remains fixed with the given set of contacts and all the other configurations are occasional or incidental. This is a property that one would expect of heterochromatin, the densely packed regions of chromatin where very little expression occurs. This may be regions where in a particular part of the life cycle of the cell, the chromatin is not used, such as developmental genes in an adult, or it can be regions of the chromatin that are only expressed when the cell is under stress. We would expect that a very prominent structure is indicative of very little dynamics; i.e., tightly packed. When a few very similar structures dominate the distribution, we would assume from this ansatz that the chromatin is shifting between various states but is only somewhat dynamic. When there are many diverse structures that have rather similar probabilities, then we can assume that there is very little differentiation between such the regions of chromatin. This latter condition would be what is characteristic of euchromatin, regions where there is generally a lot of gene expression and where the structures are open and ready for transcription.

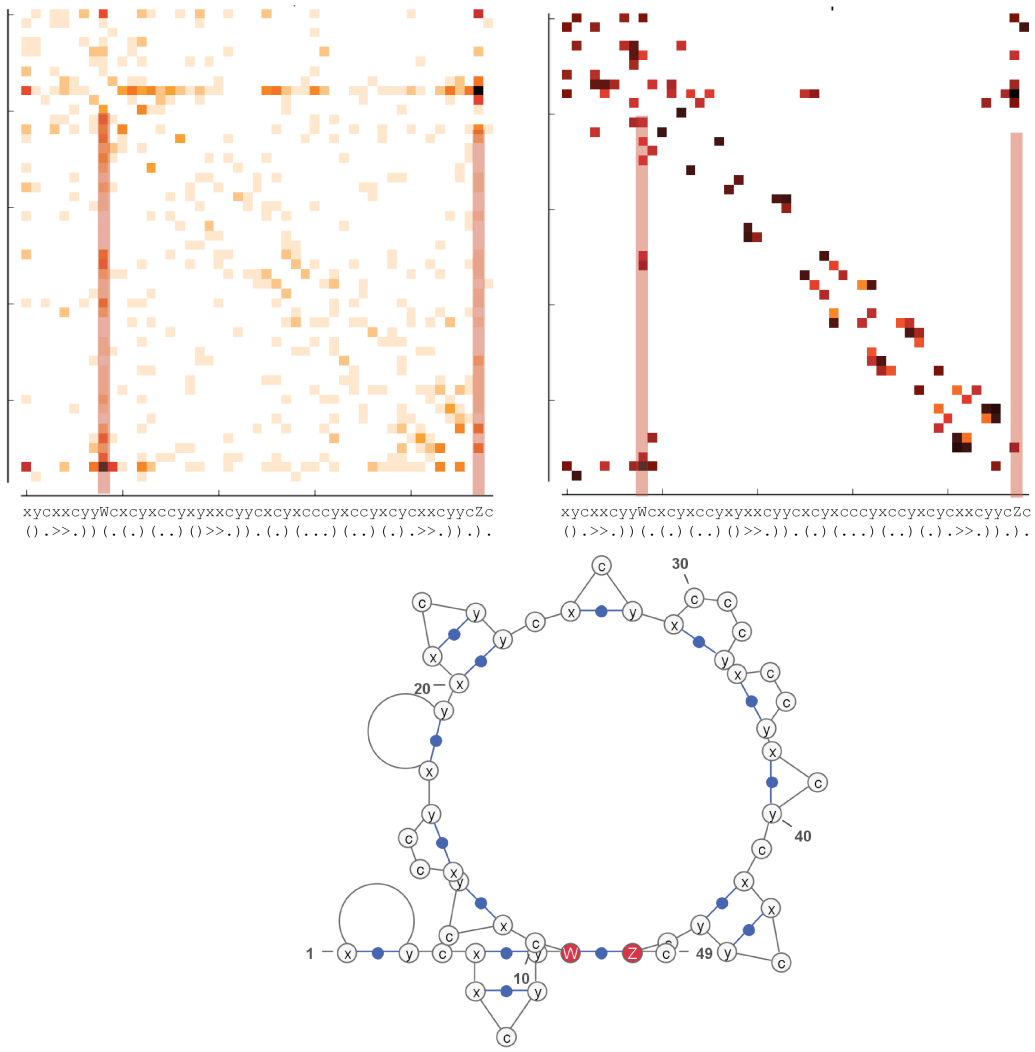

Figure 2. An example of contacts within a loop obtained from PET clusters and the weighted thermodynamic distribution of those structures based on the assumed polymer behavior of the chromatin. (a) The original heat map of the data obtained from ChIA-PET data. (b) The analysis using of predominant structures based on the Boltzmann distribution. (c) The dominant structure found in the distribution. (d) The second most dominant structure.

Fig 2 shows an example wherein Fig 2A shows the observed contact map, Fig 2B shows the dominant contacts found with the polymer model developed here, and Figs 2C show a 2D representation of the minimum free energy structure. Many features of the original heat

map are observed in the thermodynamic distribution, indicating a reasonable correspondence between the observed frequency of contacts and the actual distribution of structures that results.

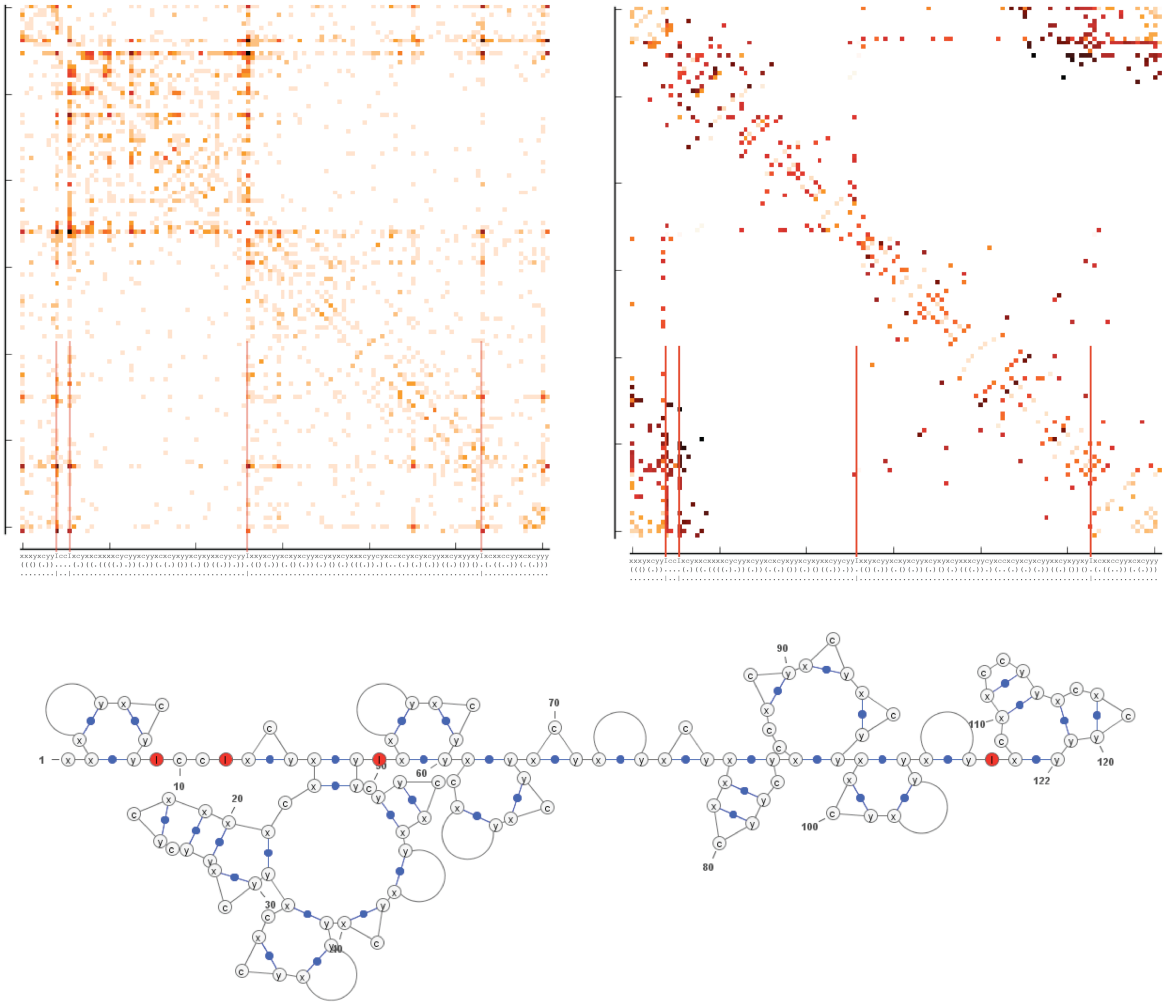

Figure 3. Result of A more complex structure in which the structure form CTCF-islands. This also shows an example of contacts within a loop obtained from PET clusters and the weighted thermodynamic distribution of those structures based on the assumed polymer behavior of the chromatin. (a) The original heat map of the data obtained from ChIA-PET data. (b) The analysis using of predominant structures based on the Boltzmann distribution. (c) The dominant structure found in the distribution. (d) The second most dominant structure.

Fig 3 is a more complex structure showing both parallel strands and CTCF islands. Fig 3A shows the observed contact map and the 1D structural notation, Fig 9B shows the dominant contacts found with the polymer model developed here with the sequence also shown for comparison, and Figs 3C shows a 2D representation of the minimum free energy structure. The 3D structure can be obtained by fitting these structures to a polymer model with restraints.

This general correspondence permits us a means to examine the condition of the chromatin in the nucleus of the cell. In this perspective, it is assumed that regions that are inactive have very stable structures – the structure with the minimum free energy (mFE) is significantly more stable (for a given genomic distance) and the distribution of alternative structures of similar free energy are few. This means there is little chance that such chromatin will be found unpacked. Likewise, if the region is active, then the mFE structure is one of many of similar probability. This would tend to be a consequence of the chromatin being dynamic, with no particular structure heavily dominating the ensemble.

Therefore, the thermodynamic probability of a given structure relative to others should be a helpful measure of to what extent that particular region of chromatin is heterochromatin or euchromatin. We were further able to show that the free energy landscape is highly predictable can and be used to identify active and inactive regions of chromatin based upon genomic distance and free energy (Figs S7 through S10).

Conclusion

We have introduced a computational algorithm for estimating the structure and dynamics of chromatin loops by analyzing the thermodynamic probability of the minimum free energy structure. This permits learning the actual structure of the chromatin in terms of 2D structural motifs. Significant correlation was suggested by the tendency of known active structures to show a less stable mFE, indicating that the observed structure of chromatin within the ensemble should be changing regularly with a high probability; likewise, the

inactive regions tended to show structures with a very large mFE with respect to genomic distance. This algorithm may help serve as an aid in determining the relative activity of the chromatin based upon the stability of the structure.

Acknowledgements

DP, WD are supported by grants from the Polish National Science Centre (2014/15/B/ST6/05082 and 2013/09/B/NZ2/00121), the European Cooperation in Science and Technology action (COST BM1405 and BM1408). DP is supported by funds from National Leading Research Centre in Bialystok and the European Union under the European Social Fund. All authors were supported by grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation" within 4DNucleome NIH program.

References

1. Szalaj P, Tang Z, Michalski P, Pietal M, Luo OJ, et al. (2016) 3D-NOME: an integrated 3-Dimensional NucleOme Modeling Engine for data-driven simulation of spatial genome organization.
2. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, et al. (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell 163: 1611-1627.
3. Dekker J, Mirny L (2016) The 3D genome as moderator of chromosomal communication. Cell 164: 1110-1121.
4. Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, et al. (2016) Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res 26: 70-84.
5. He S, Dunn KL, Espino PS, Drobic B, Li L, et al. (2008) Chromatin organization and nuclear microenvironments in cancer cells. J Cell Biochem 104: 2004-2015.
6. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-293.
7. Lewis A, Murrell A (2004) Genomic imprinting: CTCF protects the boundaries. Curr Biol 14: R284-286.

8. Wang S, Xu J, Zeng J (2015) Inferential modeling of 3D chromatin structure. Nuc Acids Res 43: e54.
9. Davie JR (1997) Nuclear matrix, dynamic histone acetylation and transcriptionally active chromatin. Mol Biol Rep 24: 197-207.
10. Davie JR, He S, Li L, Sekhavat A, Espino P, et al. (2008) Nuclear organization and chromatin dynamics--Sp1, Sp3 and histone deacetylases. Adv Enzyme Regul 48: 189-208.
11. Barbieri M, Fraser J, Lavitas LM, Chotalia M, Dostie J, et al. (2013) A polymer model explains the complexity of large-scale chromatin folding. Nucleus 4: 267-273.
12. Barbieri M, Scialdone A, Piccolo A, Chiariello AM, di Lanno C, et al. (2013) Polymer models of chromatin organization. Front Genet 4: 113.
13. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462: 58-64.
14. Nacheva GA, Guschin DY, Preobrazhenskaya OV, Karpov VL, Ebralidse KK, et al. (1989) Change in the pattern of histone binding to DNA upon transcriptional activation. Cell 58: 27-36.
15. Maeshima K, Hihara S, Eltsov M (2010) Chromatin structure: does the 30-nm fibre exist in vivo? Curr Opin Cell Biol 22: 291-297.
16. Joti Y, Hikima T, Nishino Y, Kamada F, Hihara S, et al. (2012) Chromosomes without a 30-nm chromatin fiber. Nucleus 3: 404-410.
17. Nishino Y, Eltsov M, Joti Y, Ito K, Takata H, et al. (2012) Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. EMBO J 31: 1644-1653.
18. Maeshima K, Rogge R, Tamura S, Joti Y, Hikima T, et al. (2016) Nucleosomal arrays self-assemble into supramolecular globular structures lacking 30-nm fibers. EMBO J 35: 1115-1132.
19. Guo Y, Xu Q, Canzio D, Shou J, Li J, et al. (2015) CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell 162: 900-910.
20. Xu C, Corces VG (2016) Towards a predictive model of chromatin 3D organization. Semin Cell Dev Biol 57: 24-30.
21. Munkel C, Eils R, Dietzel S, Zink D, Mehring C, et al. (1999) Compartmentalization of interphase chromosomes observed in simulation and experiment. J Mol Biol 285: 1053-1065.
22. Bystricky K, Heun P, Gehlen L, Langowski J, Gasser SM (2004) Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proc Natl Acad Sci U S A 101: 16495-16500.

23. Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, et al. (2016) 3D chromosome regulatory landscape of human pluripotent cells. Cell Stem Cell 18: 262-275.
24. Beagan JA, Gilgenast TG, Kim J, Plona Z, Norton HK, et al. (2016) Local genome topology can exhibit an incompletely rewired 3D-folding state during somatic cell reprogramming. Cell Stem Cell 18: 611-624.
25. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, et al. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523: 486-490.
26. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159: 1665-1680.
27. Sapojnikova N, Thorne A, Myers F, Staynov D, Crane-Robinson C (2009) The chromatin of active genes is not in a permanently open conformation. J Mol Biol 386: 290-299.
28. Tjong H, Li W, Kalhor R, Dai C, Hao S, et al. (2016) Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc Natl Acad Sci 113: E1663-1672.
29. Meluzzi D, Arya G (2013) Recovering ensembles of chromatin conformations from contact probabilities. Nuc Acids Res 41: 63-75.
30. Chiariello A, Bianco S, Piccolo A, Annunziatella C, Barbieri M, et al. (2014) Polymer models of the ogranization of chromosomes in the nucleus of cells. Complesso Universitario di Monte S. Angelo, Napoli, Italy.
31. Dawson W, Yamamoto K, Kawai G (2012) A new entropy model for RNA: part I. A critique of the standard Jacobson-Stockmayer model applied to multiple cross links. J Nucl Acids Invest 3: e3.
32. Dawson W, Yamamoto K, Shimizu K, Kawai G (2013) A new entropy model for RNA: part II. Persistence-related entropic contributions to RNA secondary structure free energy calculations. J Nucl Acids Invest 4: e2.
33. Dawson W, Takai T, Ito N, Shimizu K, Kawai G (2014) A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped? J Nucl Acids Invest 5: 2652.
34. Dawson W, Fujiwara K, Kawai G (2007) Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS One 2: 905.