Final Project
Biochem 218: Computational Molecular Biology
Mark Auburn
Dec. 13th, 2009

Recent Advances in the Determination of Three-Dimensional Chromosomal Conformation

Introduction

The completion of the human genome project and the advent of second- and third-generation sequencing technologies have reduced the uncertainties surrounding DNA sequence data for most organisms to an almost absurdly low level – we take it for granted that reference genomes are available for most commonly studied organisms, and that detecting sequence variations for individuals or particular cell types can be done at a relatively low cost.

By contrast, there are still substantial uncertainties surrounding three-dimensional structures in molecular biology. As a simple example, the number of resolved protein structures lags far behind the amount of sequence data available. As both a consequence and a companion of that relative lack of data, there is also a relative dearth of substantiated hypotheses concerning the mechanisms of three-dimensional activity in molecular biology.

One example of that dearth is the relative poverty of information concerning the organization of chromosomes in the cell nucleus. Although the structure of DNA is well-known, as is the basic nucleosome structure (the combination of DNA and the histone proteins), all the levels of understanding beyond that "micro-level" view are clouded in a fog comprised of little experimental data as well as unsubstantiated hypotheses.

And yet the problem of chromosome organization is a vital one for the understanding of the cell: folding a chromosome that can be several meters long in such a way that it is both dense enough to fit into the tiny cell nucleus, as well as open enough to permit the wide variety of interactions that are essential for gene expression, is clearly an important topic.

A recent paper in Science(1) that has received much popular press purports to illuminate that fog in two fundamental ways: by demonstrating a new technique for obtaining experimental data on the three-dimensional structure of chromosomes, and by advancing a new hypothesis about the fundamental three-dimensional layout of chromosomes.

This paper reviews some of the prior knowledge about that chromosomal organization, and some of the research that has gone into that knowledge, and then critiques the recent paper on both its experimental technique as well as on its structural hypothesis of chromosomal organization, and then suggests some directions in which it would be profitable to proceed.

Prior Status and Some Recent Research

Although the structure of DNA and the basic layout of the nucleosome appear to be well-understood, there are still outstanding questions relating to the dynamic nature of the nucleosome and what might be called the "micro-structure" of the chromatin fiber (2)(3)

Above the level of pure chromatin, there have been numerous attempts to gain knowledge on the organization of the chromosome itself, based mostly on model organisms, and attempting various experimental techniques.

The Sedat lab at UCSF(4) has pioneered various approaches to this problem, using three-dimensional microscopy, electron tomography and several variations on fluorescent imaging to gain knowledge from various angles.

One interesting approach that they created in 2004 was to use color barcode labeling with FISH (fluorescent in situ hybridization) to create a three-dimensional map of various genetic loci(5).

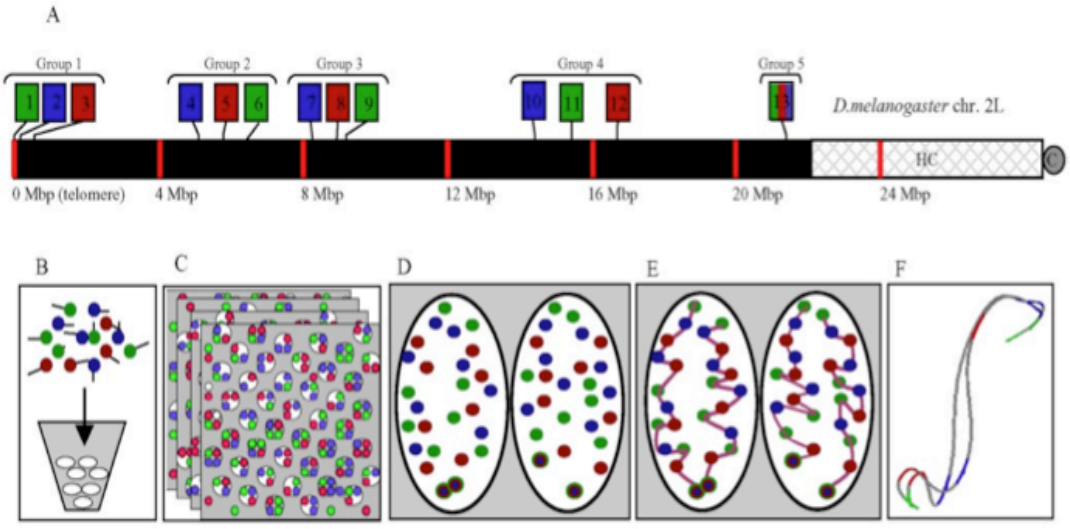


Figure 1. Experimental design. (A) A three-color, 13-probe barcode was designed which maps to *D. melanogaster* chromosome 2L. Probe localizations are based on release 3 of the Drosophila Genome Project Map. Euchromatin is shown in black (22.2 Mbp), heterochromatin is shown in gray mesh (6.2 Mbp), and the gray circle labeled C indicates the centromere. The probe labeled 13 maps to the histone locus and was labeled in all colors. The remaining probes were labeled with one color each. (B) The cut and labeled barcode probes were hybridized to cycle 14 *Drosophila* embryos. (C) Embryos were imaged using 3D wide-field fluorescence microscopy, and the data sets were deconvolved. (D) Nuclei and probe signals were segmented. (E) Chromosome paths were deduced using criteria described in the text. (F) The set of traces was subject to structural analysis.

By using three colors for a combination of 13 probes they investigated the Drosophila chromosome 2L and obtained a rough sketch of the locations of those loci, which they combined with molecular dynamics simulation and various physics-based principles to arrive at tentative conclusions for the location of those loci in the nucleus.

They confirmed that the chromosome was polarized in the Rabl orientation (i.e. that centromeres were at one nuclear pole and telomeres were at another) and that a random coil configuration described their data.

Another approach that they explored in 2007 was to use three-dimensional electron microscopy (EM) tomography to investigate reconstituted sperm nuclei of the Xenopus laevis (African clawed frog)(6)

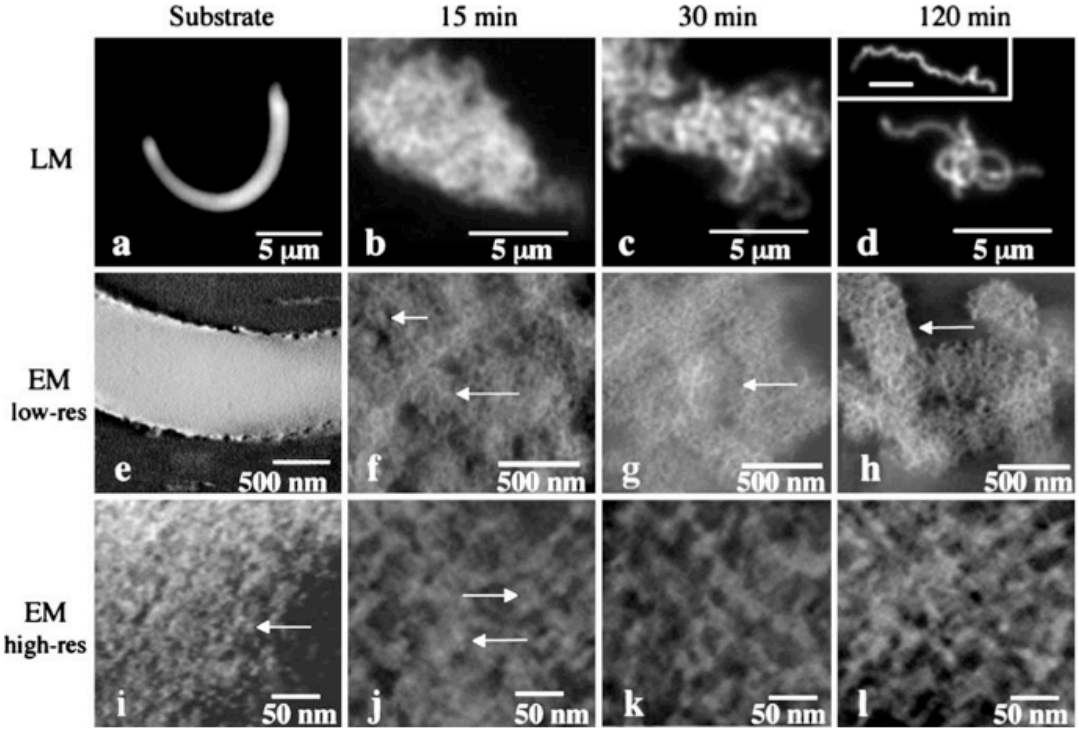


Fig. 1 Structure of *X. laevis* chromosomes at successive time points of in vitro reconstitution reaction. **a**–**d** Fluorescence microscope images of YO-PRO stained samples. The shown areas are volume projections of 1.8-μm-thick volumes. **e**–**h** Low-resolution and **i**–**l** high-resolution 3D EM reconstructions of chromosomal samples. Low-resolution images correspond to 60-nm-thick sections, and high-resolution images to 5-nm-thick sections, respectively. **a**, **e**, and **i**, *X. laevis* sperm nucleus substrate; **b**, **f**, and **j**, chromosomes after 15 min incubation; **c**, **g**, and **k**, after 30 min incubation and **d**, **h**, **l** after 120 min incubation in metaphase-arrested *X. laevis* egg extract. (*Inset*) **d** single completely resolved chromatid. **f** *Arrows* in low-resolution images mark characteristic sizes of chromatin domains in 15-min reconstitution intermediate, **g** the association of chromosome boundaries in the 30-min intermediate **h** and their resolution in the fully reconstituted chromosomes. **i** *Arrows* in the high-resolution images indicate the size of characteristic substructures in the sperm nucleus **j** and reconstituted chromatin

By creating their own atomic model of the nucleosome, and combining it with the observations from their tomographic data, they created models of the chromatin layout in three-dimensional space by applying a series of rotations and

translations to fit the data, using a variety of constraints (such as for steric hindrance and based on the observed conformations of the linker DNA between nucleosomes) to constrain the search space.

They observed a variety of interconnected clustering patterns, often in a zig-zag shape, and concluded that a random coil-like structure, as speculated elsewhere, was consistent with the data.

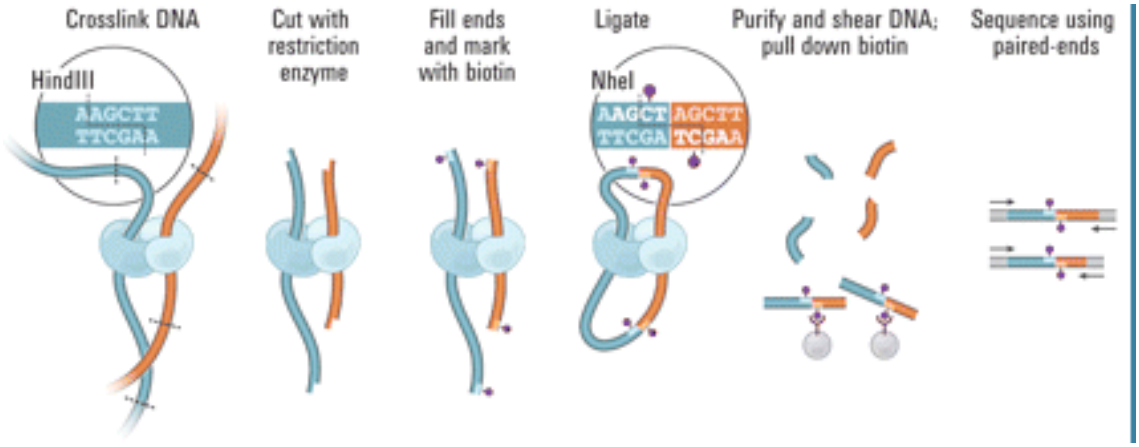
The Dekker lab(7) at the University of Massachusetts has also been active in pursuing studies of three-dimensional chromosome organization, using primarily the yeast and human genomes, particularly with aim of exploring how three-dimensional organization might affect gene regulation(8)

Chromosome conformation capture (3C), along with its cousins 4C (3C-on chip) and 5C (3C-carbon copy) have been developed by this lab as a means of gathering three-dimensional structural data for selected regions.(9) However, for these techniques it has been necessary to select the loci or regions that are to be investigated.

"Hi-C" and the Announcement of General Genomic Folding Principles

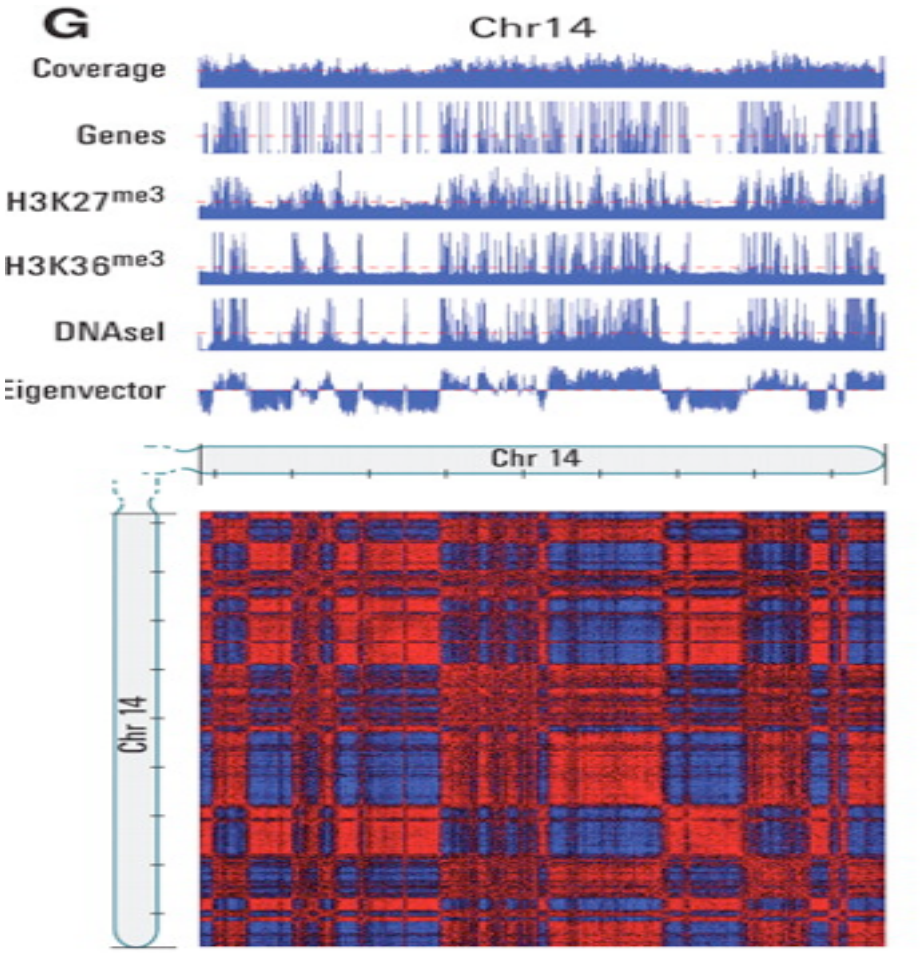
The latest research was conducted by Job Dekker of the University of Massachusetts Medical School and Eric Lander of the Broad and many coworkers, and created a new experimental technique that they dubbed "Hi-C" for determining the three-dimensional architecture of an entire genome by combining proximity-based ligation with next-generation sequencing.(1) It is a natural extension of their prior 3C/4C/5C techniques, yet it has the significant advantage of not requiring foreknowledge of which loci or regions are of interest, and allows for "whole-genome" investigation.

This experimental technique is being patented, and combines several well-known techniques together: it uses formaldehyde to link together DNA strands that are next to each other, and then cuts the strands, biotinylates them, ligates them, and then shears them and sequences them using paired-end sequencing.



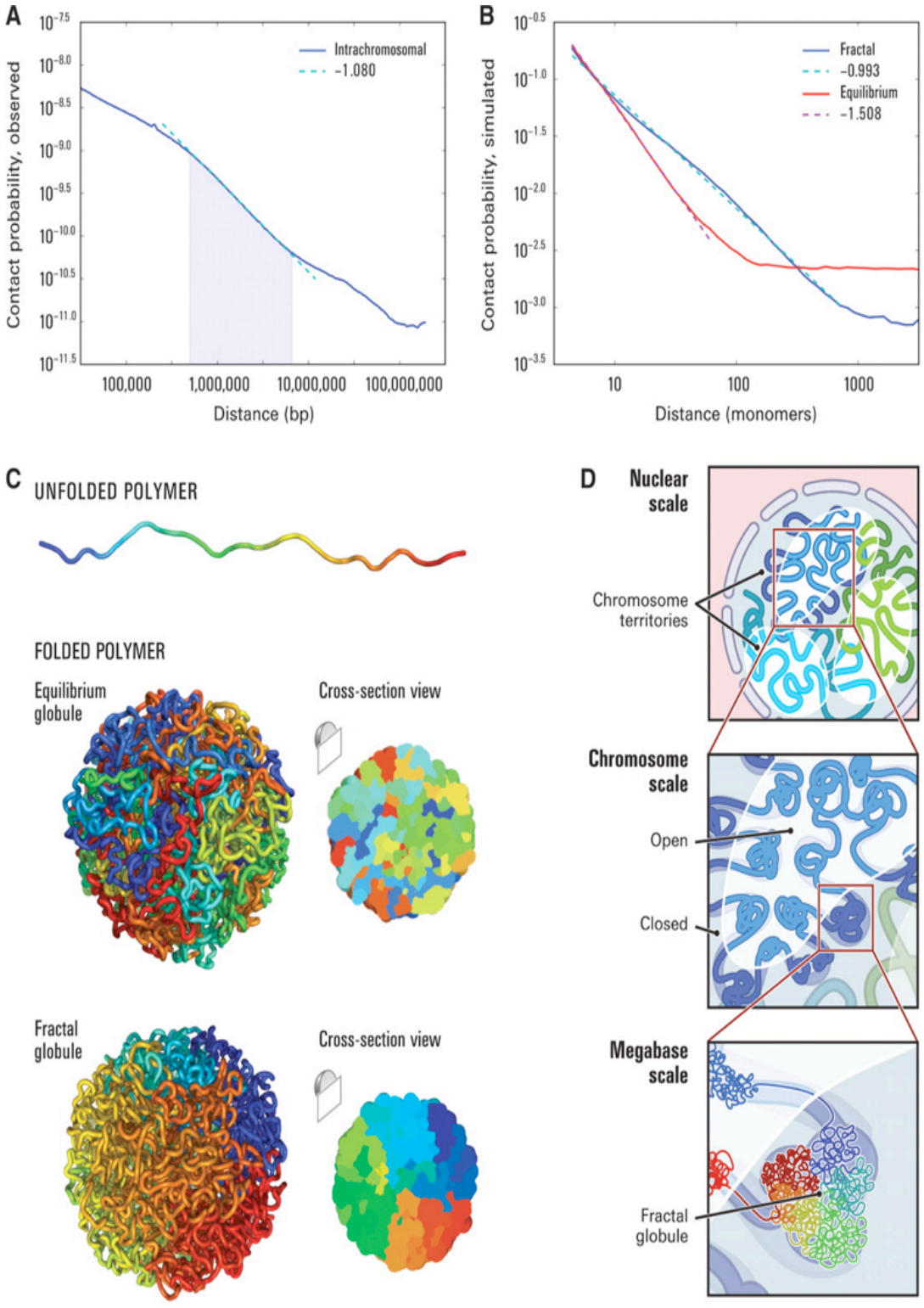
After throwing out reads that can't be uniquely matched, this techniques allow them to arrive at effectively a correlation matrix between each resolved read position and every other resolved-read position that has been cross-linked with it, allowing a series of "contact matrices" (matrices of which positions have been successfully cross-linked with each other) to be built, both across chromosomes and within chromosomes.

As tests of their technique, they repeated it using the same restriction enzyme, and also with a different restriction enzyme, and obtained substantially similar results.



Although they only profess to be able to resolve to 1Mbp at this time using this technique, that allowed them to confirm several features known to exist already in the three-dimensional chromosomal organization: the existence of chromosome territories and that certain small, gene-rich chromosomes tend to be close to each other. They also identified what they called “compartments”: spatially segregated areas of open and closed chromatin that are genome-wide compartments. In the “open” compartment, gene expression is active, and in the “closed” compartment gene expression is relatively inactive. They offer a browser of this data at (10).

In the figure above, higher-than-expected correlations are marked with red, lower-than-expected with blue, and the relationships with the distributions of genes can be seen.



In addition, they used the correlation or contact matrices to investigate the overall topological structure of the chromosomes. They concluded, after conducting numerous simulations of various three-dimensional space-filling structures, that the pattern of contact probabilities was most consistent with a so-called “fractal

globule", as opposed to a "equilibrium globule". A fractal globule is defined as a knot-free conformation that is fractally defined; in other words, it consists of globules folded together in the same way that those globules have been folded together (a fractal pattern, similar to many other fractal patterns in nature or mathematics). A fractal globule can unfold and refold with relative ease, as it has no knots to hinder its activity. By contrast, an equilibrium globule is densely knotted and unfolds only with difficulty.

These Monte Carlo simulations were of polymer configurations consistent with known properties of chromatin fiber – since double-stranded DNA has a known persistence length of about 50nm (in other words, at less than 50nm it behaves like a rigid rod, and at greater than 50nm it behaves like a worm-like chain). For each conformation that they created, they investigated the "Alexander polynomial" for that conformation, which is a metric of the amount of knotting created in that conformation (in knot theory, the Alexander polynomial is a so-called "knot invariant", which means that it is constant regardless of twistings and stretchings, as long as the knots themselves aren't cut). One version of their tool can be referenced at (11)

Based on the slope of the contact probability curve, they concluded that a fractal globule fit the data better than an equilibrium globule.

Questions and Further Directions

This new research raises numerous questions and creates many possibilities for further research.

1. Relatively little data is offered that the new Hi-C technique is reproducible (against itself) and valid (in comparison to past chromosome conformation techniques, in particular 3C/4C/5C). It would be valuable to conduct additional tests, both with these same cell types and other cell types, to validate reproducibility and validity.
2. An interesting combination of past and current research would be to combine Hi-C and FISH loci testing in some manner. Hi-C excels in showing correlations, but fails to show actual three-dimensional location (particularly in regards to the nucleus). FISH loci testing, especially as with the FISH bar-coding paper above, fails to include more than a few loci but shows physical location. If a few FISH loci could be included in a Hi-C test (if that technique is possible), it would be possible to make interesting inferences about the physical layout of the actual chromosomes against the nucleus and each other.
3. One interesting test of these Hi-C correlations would be to test the relationship between promoters and enhancers, and other regulatory elements, especially for enhancers that are significantly far away from the promoter, and to see if their three-dimensional distance (or correlation) is

significantly less than their sequence distance. That would be an interesting test that would, if present, offer a measure of reassurance.

4. Longitudinal studies of Hi-C correlations over different cell types, and over time, would be interesting, in that it would offer interesting comparisons against existing microarray experiments of gene expressivity with various cell types and developmental stages, and would offer further confirmation of the hypothesis that "open compartments" correlate with gene expressivity.
5. One disappointment in this research is that the Monte Carlo simulations to investigate the globule nature were based not on the actual correlations, as found in the data, but on the summary statistics derived from the data. It would have been more valid, if feasible at the 1Mbp range, to have conducted simulations based on the actual contacts that were found, along with the physical assumptions that they used. In particular, it is intriguing to speculate on whether the density of the globule, as well as whether it enjoys fractal or equilibrium characteristics, varies in particular regions, for it is most unlikely that it is purely fractal or purely equilibrium, and using summary statistics hides that fact.
6. In particular, one way to conduct the simulations that would have resulted in more interesting results would have been to conduct "fractal" or "partial" tests of their data. By taking successive fractions of their correlation data, it would have been possible to test whether each fraction was in itself fractal or whether there was intrastructural variation in character.
7. Whether the three-dimensional chromosome organization is fractal or equilibrium or of some other nature, it would be an interesting area for research to investigate how that nature is created and maintained, and what the possible breakdowns are for the folding and unfolding mechanisms. The earlier paper(6) had a long discussion of mechanisms for the creation and maintenance of chromosome organization; it would be most intriguing to extend that discussion to include fractal globule character.

References

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009 Oct;326(5950):289-293.

2. van Holde K. Scanning Chromatin: a New Paradigm? Journal of Biological Chemistry 2006;281(18):12197-12200.

3. Vanholde K, Zlatanova J. Chromatin fiber structure: Where is the problem now? Seminars in Cell & Developmental Biology 2007;18(5):651-658.

4. Research [Internet]. [date unknown];[cited 2009 Dec 14] Available from: http://msg.ucsf.edu/sedat//research.html

5. Lowenstein MG, Goddard TD, Sedat JW. Long-range interphase chromosome organization in Drosophila: a study using color barcoded fluorescence in situ hybridization and structural clustering analysis. Mol. Biol. Cell 2004 Dec;15(12):5678-5692.

6. König P, Braunfeld MB, Sedat JW, Agard DA. The three-dimensional structure of in vitro reconstituted Xenopus laevis chromosomes by EM tomography. Chromosoma 2007 Aug;116(4):349-372.

7. Dekker Lab : Welcome [Internet]. [date unknown];[cited 2009 Dec 14] Available from: http://my5c.umassmed.edu/welcome/welcome.php

8. Dekker J. Gene Regulation in the Third Dimension. Science 2008;319(5871):1793-1794.

9. Lajoie BR, van Berkum NL, Sanyal A, Dekker J. My5C: web tools for chromosome conformation capture studies. Nat Meth 2009;6(10):690-691.

10. Hi-C Data Browser [Internet]. [date unknown];[cited 2009 Dec 14] Available from: http://hic.umassmed.edu/heatmap/heatmap.php

11. Knots in the proteins - prediction server [Internet]. [date unknown];[cited 2009 Dec 14] Available from: http://knots.mit.edu/