

# *Structure determination of genomic domains by satisfaction of spatial restraints*

## **Chromosome Research**

Cytogenetics, genomics,  
chromatin and the nucleus

ISSN 0967-3849

Volume 19

Number 1

Chromosome Res (2010)

19:25-35

DOI 10.1007/

s10577-010-9167-2



**Your article is protected by copyright and all rights are held exclusively by Springer Science+Business Media B.V.. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your work, please use the accepted author's version for posting to your own website or your institution's repository. You may further deposit the accepted author's version on a funder's repository at a funder's request, provided it is not made publicly available until 12 months after publication.**

# Structure determination of genomic domains by satisfaction of spatial restraints

Davide Baù · Marc A. Marti-Renom

Published online: 29 December 2010  
© Springer Science+Business Media B.V. 2010

**Abstract** The three-dimensional (3D) architecture of a genome is non-random and known to facilitate the spatial colocalization of regulatory elements with the genes they regulate. Determining the 3D structure of a genome may therefore probe an essential step in characterizing how genes are regulated. Currently, there are several experimental and theoretical approaches that aim at determining the 3D structure of genomes and genomic domains; however, approaches integrating experiments and computation to identify the most likely 3D folding of a genome at medium to high resolutions have not been widely explored. Here, we review existing methodologies and propose that the integrative modeling platform (<http://www.integrativemodeling.org>), a computational package developed for structurally characterizing protein assemblies, could be used for integrating diverse experimental data towards the determination of the 3D architecture of genomic domains and entire genomes at unprecedented resolution. Our approach, through the visualization of looping interactions between distal regulatory elements, will allow for the characterization of global chromatin features and their relation to gene expression. We illustrate our work by outlining the recent determination of the 3D architecture of the  $\alpha$ -globin domain in the human genome.

**Keywords** chromosome conformation · chromatin structure · integrative modeling · computational structural biology

## Abbreviations

1D	One-dimensional
3D	Three-dimensional
3C	Chromosome conformation capture
4C	Circular chromosome conformation capture
5C	Chromosome conformation capture carbon copy
Bp, Kb, Mb	Base pairs, kilo base pairs, mega base pairs
ChIP	Chromatin immunoprecipitation
DamID	DNA adenine methyltransferase identification
FISH	Fluorescence in situ hybridization
IMP	Integrative modeling platform
LMA	Ligation-mediated amplification
nm	Nanometer
NMR	Nuclear magnetic resonance
PCR	Polymerase chain reaction
RMSD	Root mean square deviation

## Introduction

The three-dimensional (3D) architecture of a genome facilitates the colocalization in space of sequentially distant loci, which is essential for carrying out their specific functions (Takizawa et al. 2008). The

D. Baù · M. A. Marti-Renom (✉)  
Structural Genomics Unit, Bioinformatics and Genomics  
Department, Centro de Investigación Príncipe Felipe,  
Av. Autopista del Saler, 16,  
46012 Valencia, Spain  
e-mail: mmarti@cipf.es

determination of the 3D architecture of whole genomes or genomic domains is deemed necessary for characterizing how genes are regulated. Unfortunately, existing light microscopy technologies like fluorescence in situ hybridization (FISH), although very informative, are not yet sufficient to provide high-resolution information on chromatin loops and the interactions between distal DNA segments. Therefore, an integrative and general approach for determining the spatial organization of chromatin may prove very useful not only for identifying long-range relationships between genes and distant regulatory elements but also for elucidating chromatin higher-order folding principles.

Previously, chromatin conformation has been modeled, for example, using polymer physics (Dekker et al. 2002; Mateos-Langerak et al. 2009) and molecular dynamics (Wedemann and Langowski 2002). Such methods rely on physics-based approaches and only partially leverage the current wealth of experimental data on chromatin folding; however, they have proven to be very valuable for understanding general features of chromatin fibers, including flexibility, compaction, and unpacking (Dekker 2008; Wachsmuth et al. 2008; Langowski and Heermann 2007; Rosa and Everaers 2008). This special issue of the *Chromosome Research* journal includes several such approaches (Fritsch and Langowski 2010; Mirny 2010).

Higher resolution experimental techniques such as chromosome conformation capture (3C)-based approaches (Dekker et al. 2002; Lieberman-Aiden et al. 2009; Dostie et al. 2007; Zhao et al. 2006; Simonis et al. 2006) have prompted the development of new hybrid methods that aim at integrating experiments and computation for identifying the most likely 3D folding of a chromatin domain at medium to high resolutions. The 3C-driven approaches, in combination with computational modeling, have so far been capable of generating medium resolution models of the topological conformation of the *HoxA* cluster (Fraser et al. 2009), the  $\alpha$ -globin domain (Baù et al. 2010), and the yeast genome (Duan et al. 2010).

This review starts by describing existing 3C-based methodologies for 3D determination of genomic domains and proposes that the integrative modeling platform (IMP, <http://www.integrativemodeling.org>) (Alber et al. 2007a) can be used for integrating such experimental data towards the determination of their 3D architecture at unprecedented resolutions. We then

review the modeling of the 3D architecture of the  $\alpha$ -globin domain in chromosome 16 of the human genome, which has been recently described in detail (Baù et al. 2010). This review ends by summarizing our thoughts on future directions for the structure determination of genomic domains.

### Chromatin interaction maps for 3D modeling

3C-based methods allow the investigation of the overall spatial organization of genomic domains, chromosomes, and entire genomes (Miele and Dekker 2009). Briefly, 3C-based methods rely on formaldehyde cross-linking between spatially close DNA loci through their corresponding bound proteins. Cross-linked DNA is then digested with a specific restriction enzyme and ligated under very diluted conditions that favor intramolecular (i.e., the cross-linked fragments) over intermolecular ligation. Cross-linking reversal and ligation product quantification by the polymerase chain reaction (PCR) using locus-specific primers returns the frequencies at which interactions occur. Given that the original 3C technique (Dekker et al. 2002) was only applicable to single pairwise loci at a time and within a relatively small DNA region (up to hundreds of kilo base pairs), there has been a plethora of new approaches that expanded 3C. For example, the so-called circular chromosome conformation capture (4C) techniques allow for the characterization of a genomic domain by a one-to-many loci interaction analysis. Those approaches take advantage of the fact that most of the 3C ligation products are already circular or can be easily circularized and then inversely amplified by PCR. Four different laboratories developed 4C technologies in parallel, which differ in the restriction enzymes used, the step at which the circular DNA is formed and the analysis of the amplified fragments. Such methods include circular 3C (Zhao et al. 2006), 3C-on-chip (Simonis et al. 2006), open-ended 3C (Wurtele and Chartrand 2006), and “olfactory receptor” 3C (Lomvardas et al. 2006). More recently, the 3C carbon copy technology (5C) was developed to allow the simultaneous and parallel detection of interactions within relatively large genomic domains or even entire chromosomes (Dostie and Dekker 2007; Dostie et al. 2006). In 5C, the PCR step of 3C is replaced by ligation-mediated amplification (LMA) followed by the detection of



ligation products. With LMA, it is possible to use simultaneously thousands of primers, allowing the parallel detection of millions of chromatin interactions. Thus, the generated 5C library is an “amplified” version of the 3C library that can be analyzed by microarray analysis or deep sequencing. Since 5C experiments are designed so that 5C primers anneal across the 3C ligation products, the specific design of the 5C primers allows interrogating multiple pairwise loci interactions in an all-against-all fashion (Lajoie et al. 2009). Such experiments are thus suitable for generating complete and extensive loci interaction matrices for large genomic regions (up to a few mega base pairs). Finally, the Hi-C technology, which allows for an unbiased genome-wide analysis, was recently developed to overcome the need for predefining a set of target loci to investigate (van Berkum et al. 2010). Hi-C’s key step is the imprinting of ligated products (i.e., products of interacting loci) with a biotin marker that later on is precipitated with streptavidin beads. Such a step allows for specifically rescuing ligation products and discarding non-ligated DNA, which is necessary for further large-scale genome-wide sequencing of the ligation products. Hi-C was recently applied to the entire human genome at 1 Mb resolution (Lieberman-Aiden et al. 2009).

The 3C-based methods result in a measure of the frequency of interaction between loci located within the studied genomic domain; however, they do not give direct information on the spatial distances between the interacting loci. It is in this scenario that the integration of 3C-based experiments with computational analysis is necessary for further determining the 3D conformation of a genomic domain. For example, Dostie and colleagues developed a suite of computer programs to identify the so-called “chromatin conformation signatures” using 5C data (Fraser et al. 2009). The work resulted in distinct structures of the *HoxA* cluster depending on the cellular differentiation stage. Starting from a random conformation, the models were iteratively changed to obtain a 3D conformation that minimized its root mean square deviation to the theoretical interloci distances calculated as simply the inverse of the 5C interaction frequencies. The final models were then used to visualize the chromatin conformation signatures of the *HoxA* cluster. More recently, Noble and colleagues

(Duan et al. 2010), using a similar approach, built 3D models of the entire yeast genome coupling 4C (Simonis et al. 2006) with massively parallel sequencing. Interaction frequencies calculated by 4C were converted into distances upon the assumption that polymer packing determines intrachromosomal distances (Bystricky et al. 2004). In particular, interaction frequencies were translated into nuclear distances by assigning 130 bp of packed chromatin to a length of 1 nm. Each chromosome was then represented as a series of beads of 10 Kb which were assigned to the closest restriction enzyme fragment resulting from the 4C experiment. Finally, a 3D model of the entire yeast genome were constructed by minimizing an objective function that scored the fitting of all bead distances to the theoretical distances as derived from the 4C experiments.

As outlined above, 3C-based experimental data about the structure of genomic domains can only be translated into a 3D model via computational methods. Next, we propose that the IMP modeling software can be used for such a task.

### Structure determination by IMP

IMP’s conceptual framework for structure determination is similar to the determination of protein structure by two-dimensional (2D) nuclear magnetic resonance (NMR) spectroscopy, where the nuclear Overhauser effect between nuclear spins is used to observe, via the 2D-nuclear Overhauser effect spectroscopy (NOESY) spectra, correlations between resonances from spins that are spatially close. In NMR, a polypeptide is represented by its atoms (particles in IMP), which will be placed in the 3D space based on the spatial distances between them (restraints in IMP) calculated from the 2D-NOESY maps (Wagner et al. 1987). In contrast to constraints, restraints are subject to probability distributions allowing the restrained particles to move within predefined limits. The final 3D ensemble of solutions of a biomolecule corresponds to the spatial positioning of all atoms that best satisfies the input experimental restraints. In contrast to NMR determination, which relies on 2D-NOESY data, IMP was developed as a general platform for simultaneously integrating diverse structural information available about the object of interest (Alber et al. 2008). Such data may greatly vary in accuracy and

resolution and may originate from any type of experimental or theoretical observations of the system. Data integration by IMP normally results in a deterministic ensemble of solutions of higher resolution than any of the individual observations (Alber et al. 2007b). The IMP conceptual framework consists of four steps: representation, scoring, optimization, and analysis.

**Representation** The first step in the integration of experimental data into a computational framework is the definition of an adequate representation of the system so that the use of the available information makes the search of the 3D conformational space feasible. Indeed, the detail of representation (or resolution) of the system determines the accessible conformational space. In other words, coarse-grained representations are more suitable for large conformational searches while fine-grained representations require more computational power to explore the same search space. IMP represents an object by a set of hierarchical particles and their properties (including their Cartesian coordinates that spatially position them). The hierarchy of the particles allows for a flexible representation of the system at different resolutions, which allows for the appropriate representation of the diverse input data.

**Scoring** The key step in structure determination by IMP is the proper evaluation of the generated models (i.e., the different solutions compatible with the input data). Therefore, the observations about the system—be it experimental, physical, or statistical—need to be translated into measurable and formulable relationships between the particles that represent the system. For this purpose, IMP uses joint probability density functions (pdf) affecting attributes of the particles including their Cartesian coordinates. Each independent observation of the system thus results in a number of pdfs affecting one or many particles. The final scoring function, also called the IMP objective function, will be then the sum of the individual pdfs affecting all particles in the system. The functional forms of the restraints implemented in IMP are diverse and were initially developed to determine the structure of protein assemblies (Alber et al. 2008).

**Optimization** Once the system is represented at the appropriate scale/s and the relationship between the particles is formulated based on the observations, a

conformational solution of the modeled object is obtained by minimizing the IMP objective function. That is, simultaneously reducing the violations of all imposed restraints. Since many different conformational solutions could satisfy (to a certain degree) the imposed restraints, it is necessary to generate a large number of independent structures to ensure an adequate conformational search. The selection of the optimization protocol in IMP depends on the representation and scoring schema of the system (Alber et al. 2008).

**Analysis** Finally, the structural analysis of the resulting ensemble of possible solutions consistent with the input restraints will reveal important aspects of the IMP modeling. It may inform, for example, on the degree of satisfaction of the imposed restraints, conflicts between different experimental observations or the intrinsic structural variability of the modeled object. Moreover, such analysis may prove useful for designing new, more informative experiments which may help to increase the resolution of the models.

Next, we briefly describe our recent work (Baù et al. 2010) using IMP for determining the 3D conformation of the  $\alpha$ -globin domain in the human chromosome 16 based solely on 5C experimental observations.

### 3D models of the $\alpha$ -globin domain using 5C data: a proof of principle

The  $\alpha$ -globin domain which is located in human chromosome 16 (Fig. 1a), is a model to study the mechanisms of long-range gene regulation (Hughes et al. 2005; Higgs et al. 2007; Higgs and Wood 2008). The 5C experiments were used to obtain a comprehensive interaction map of the  $\alpha$ -globin locus over an ~500-Kb region, which corresponded to the ENm008 ENCODE pilot region (Birney et al. 2007) was obtained by 5C experiments. The 5C primers were designed at *Hind*III sites using the My5C server (Lajoie et al. 2009). In total, 30 forward primers and 25 reverse primers were designed throughout the ~500-Kb region with the capacity for detecting 750 pairwise chromatin interactions. The 5C experiments were performed in  $\alpha$ -globin expressing K562 cells where long-range interactions are expected to occur between the  $\alpha$ -globin genes and their regulatory hypersensitive elements located upstream. The resulting 2D frequency matrix for interactions

between loci along the studied region was transformed by means of  $Z$  scoring. First, a  $\log_{10}$  transformation of the raw frequencies normalized the interaction 5C matrix. Second,  $Z$  scores of the  $\log_{10}$  values for interacting fragments  $i$  and  $j$  were computed as:

$$Z \text{ score}_{i,j} = \frac{(\mu - f_{i,j})}{\sigma}$$

where  $f_{i,j}$  was the  $\log_{10}$  5C frequency between fragments  $i$  and  $j$ , and  $\mu$  and  $\sigma$  were the average and standard deviation of the  $\log_{10}$  frequencies of the whole 5C matrix (Fig. 1b).

**Representation** Modeling the 3D structure of a complex system such as a genomic domain implies the exploration of a large conformational space. Determining the exact level of representation is important for balancing the need to capture all observations about the system and the required computational time to generate solutions that satisfy such observations. The  $\alpha$ -globin domain was represented by a set of 70 particles, one for each of the resulting restriction fragments after digestion by *HindIII*. Each particle in the system was assigned an excluded volume defined as a sphere of radius proportional to the particle's size (in base pairs). Considering the canonical 30-nm fiber, the relationship between length and base content was set to 0.01 nm per base pair (Gerchman and Ramakrishnan 1987). After the system was properly represented, a set of restraints was assigned to each particle defining their relative 3D position and thus the final spatial organization of the whole  $\alpha$ -globin domain.

**Scoring** The 5C data consist of a matrix of interaction frequencies between restriction fragments which do not give a direct measure of the Euclidean distances between the particles representing the fragments. Nevertheless, given the assumption that spatially adjacent fragments interact more frequently than spatially distant ones, the frequency by which two loci are captured in an experiment represents the probability for those two loci to be spatially close in a given cell state (Lieberman-Aiden et al. 2009). Restraints were assigned to each of the 70 particles in our system following three basic assumptions: (1) neighbor (i.e.,  $i$  to  $i+1..2$ ) and non-neighbor (i.e.,  $i$  to

$i+3..n$ ) particles followed different 5C  $Z$  scores distribution which reflected their different response in 5C experiments (Dekker 2006); (2) consecutive particles were spatially restrained proportionally to the occupancy of their chromatin fragments with a relationship of 0.01 nm/bp, assuming a canonical 30-nm fiber; and (3) two non-neighbor fragments could not get closer than 30 nm, which corresponds to the diameter of the chromatin fiber. Given these assumptions, we were able to define two different linear relationships to map 5C  $Z$  scores onto Euclidean distances that were then set to restraint pairs of particles. Distances between neighbor particles were derived from the relationship between the sum of the radii of the experimental restriction enzyme fragments involved in the interaction and their corresponding 5C  $Z$  scores (red points and regression line in Fig. 1c). For non-neighbor particles, three IMP parameters were empirically determined to define the type and magnitude of the restraint to be applied: the minimum possible distance between two non-interacting particles (yellow regression line in Fig. 1c) and a  $Z$  score upper- and lower-bound cutoffs (blue dashed vertical lines in Fig. 1c). The optimal values for these three IMP parameters were obtained by maximizing the correlation coefficient between the input 5C frequency matrix and a contact map generated from the 3D models built by IMP. The correlation coefficient was 0.69 for the  $\alpha$ -globin models built using 400 nm as the maximum distance between non-neighbor fragments,  $-0.1$  for the lower-bound  $Z$  score cutoff, and  $+0.9$  for the upper-bound  $Z$  score cutoff (Fig. 1c). Three different types of restraints were then applied to the 70 particles: (1) harmonic restraints, which were set between pairs of neighbor particles and between pairs of non-neighbor particles with  $Z$  scores higher than the upper-bound cutoff, maintained two particles at a given equilibrium distance (red in Fig. 1d); (2) lower-bound harmonic restraints, which were set between pairs of non-neighbor particles with  $Z$  scores lower than the lower-bound cutoff, maintained two particles farther than a given equilibrium distance (blue in Fig. 1d); and (3) upper-bound harmonic restraints, which were set between pairs of neighbor particles with no experimental data available, maintained two particles within a given equilibrium distance (green in Fig. 1d). For example, an upper-bound harmonic was set between neighbor particles 26 and 28 with missing 5C data, a pair of harmonic restraints was set between non-neighbor

particle 20 and particles 26 and 28 with 5C Z scores higher than +0.9, and a pair of lower-bound harmonic restraints was set between non-neighbor particle 62 and particles 26 and 28 with Z scores lower than -0.1 (Fig. 1e). In total, the 70 particles representing the  $\alpha$ -globin domain were restrained by 1,049 restraints, of which 235 were harmonic, 709 lower-bound harmonic, and 105 upper-bound harmonic.

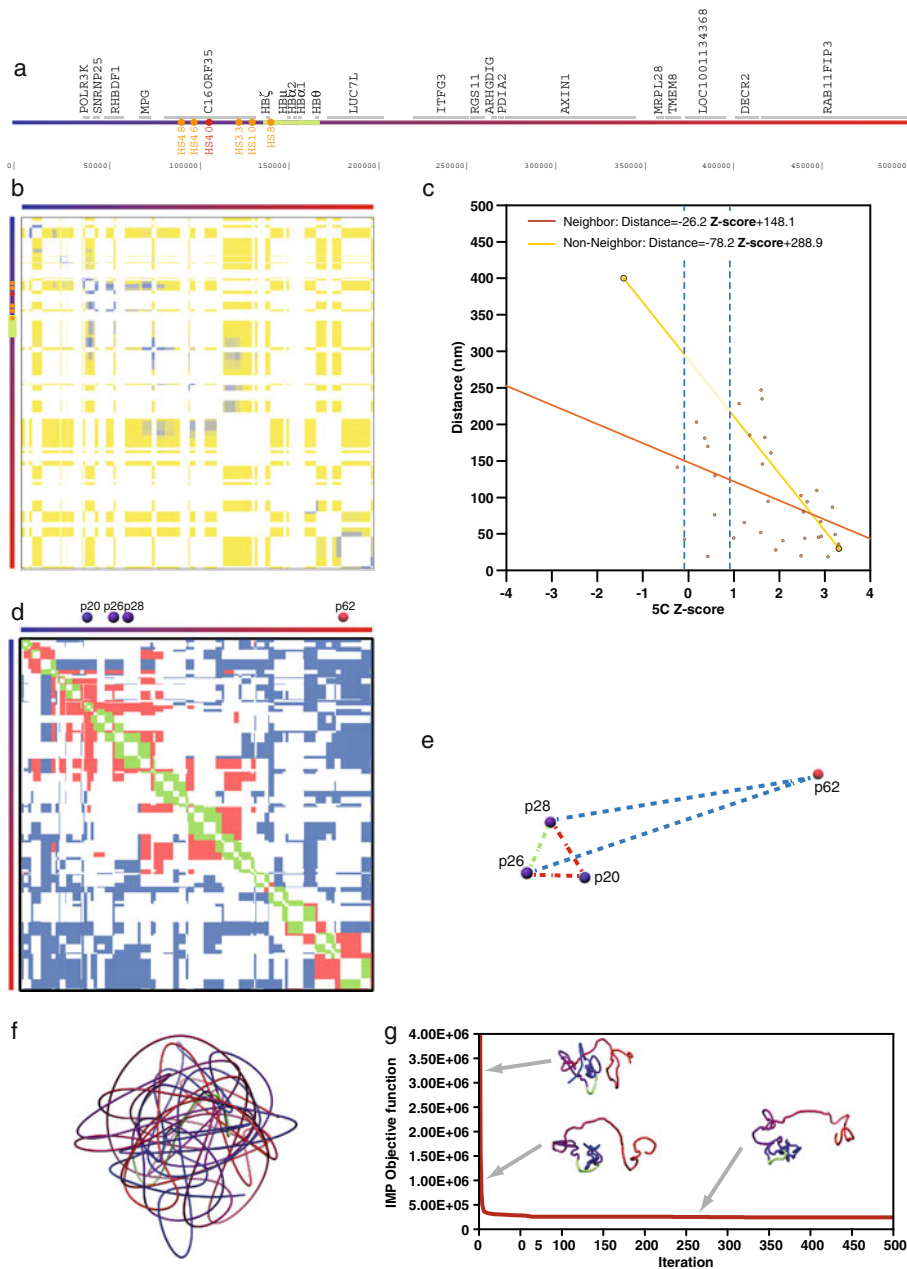
**Optimization** Once the system is represented by a set of particles and their imposed restraints, IMP 3D structure determination is expressed as an optimization problem. Starting from a random configuration (Fig. 1f), particles are moved in search for their relative position in space (i.e., a 3D conformation) that minimally violates the imposed restraints (Fig. 1g). The quality of a model is then measured by the IMP objective function, which is the sum of violations of all individual restraints applied to the particles representing the system. Each restraint is scored based on the difference between the distance in the model and its equilibrium distance. For harmonic restraints, the score scales with the square of the distance difference. When all the restraints in the final model are consistent with the input experimental data, the IMP objective function of the system is 0, whereas inconsistencies are penalized depending on the magnitude of the individual violations. The optimization protocol adopted in our method consisted of 500 Monte Carlo rounds and five local optimization steps taken at each round with a simulated annealing protocol (Kirkpatrick et al. 1983). At each Monte Carlo step, a defined set of particles was moved by translating their Cartesian coordinates limited by a defined Gaussian distribution with a sigma of 0.25 and centered in their current position. According to the Metropolis criteria, a move was accepted with a probability proportional to the difference in the IMP objective function before and after the move and the temperature of the system. To warrant proper search of the conformational space, this protocol was run for 50,000 times with different random starting conditions. Therefore, our model building by IMP resulted in 50,000 different solutions of the  $\alpha$ -globin domain structure.

**Ensemble analysis** The 10,000 solutions with the lowest IMP objective function were selected out of

**Fig. 1** Structure determination of the  $\alpha$ -globin domain. **a** 1D Map of 0.5 Mb ENm008, including the  $\zeta$ ,  $\mu$ ,  $\alpha 2$ ,  $\alpha 1$ , and  $\theta$  globin genes. Gray lines above the linear representation colored from blue (telomere) to red (centromere) show annotated genes in the region. Red and orange circles localize the HS40 and other  $\alpha$ -globin-related HS sites. **b** 5C Z scores matrix. Blue to yellow color indicate positive to negative Z scores. White color indicates non-interrogated pairs. For easy inspection, the 1D representation of the ENm008 is used in the x- and y-axes of the plot. **c** IMP calibration to assess the linear relationship between 5C Z scores and equilibrium distances between neighbor (red linear fitting) and non-neighbor fragments (yellow line). Two vertical dashed blue lines indicate optimal Z scores upper- and lower- cutoffs. **d** Harmonic (red), lower-bound harmonic (blue) and upper-bound harmonic (green) restraints applied to pairs of restrained fragments during simulation. For easy inspection, the 1D representation of the ENm008 is used in the x- and y-axes of the plot. **e** Schematic representation of four particles (numbers 20, 26, 28, and 62) and the applied restraints between them, which are represented as a dashed line colored as the restraint type. For easy inspection, the particles 1D position is also shown in panel d. **f** Starting structure with random positioning of all 70 particles within the  $\alpha$ -globin domain. **g** Schematic representation of a typical optimization process for a single simulation. An optimal configuration is achieved after the minimization of the IMP objective function summing all violated restraints. Images of the structures shown were generated using the Chimera program (Pettersen et al. 2004)

the 50,000 simulations and used in an all-against-all structure comparison by rigid-body superposition. The resulting structure comparison matrix was then input into the Markov Cluster Algorithm program (Enright et al. 2002) to generate unsupervised sets of clusters of similar structures. The 10,000 solutions resulted in 393 clusters of superposed solutions with the top 10 largest clusters accounting for 26% of the 10,000 solutions (Fig. 2a). Structural clusters number 1 and 2 resulted in the most populated clusters (483 and 314 solutions, respectively) and had the lowest IMP objective function scores. This result indicated that for most of the simulations, we identified the same final 3D conformation (i.e., a global minima of the IMP objective function). The overall features of the different structures in a given cluster were well conserved and always resulted in a two-globular structure of the  $\alpha$ -globin domain (Fig. 2b). The observed differences between clusters likely reflected the high variability of chromatin conformation of the  $\alpha$ -globin domain in K562 cells, which could be related to the high transcription levels of the resident genes or to the chromosome structural variability in a cancer cell line such K562. Thus, the ensemble of all

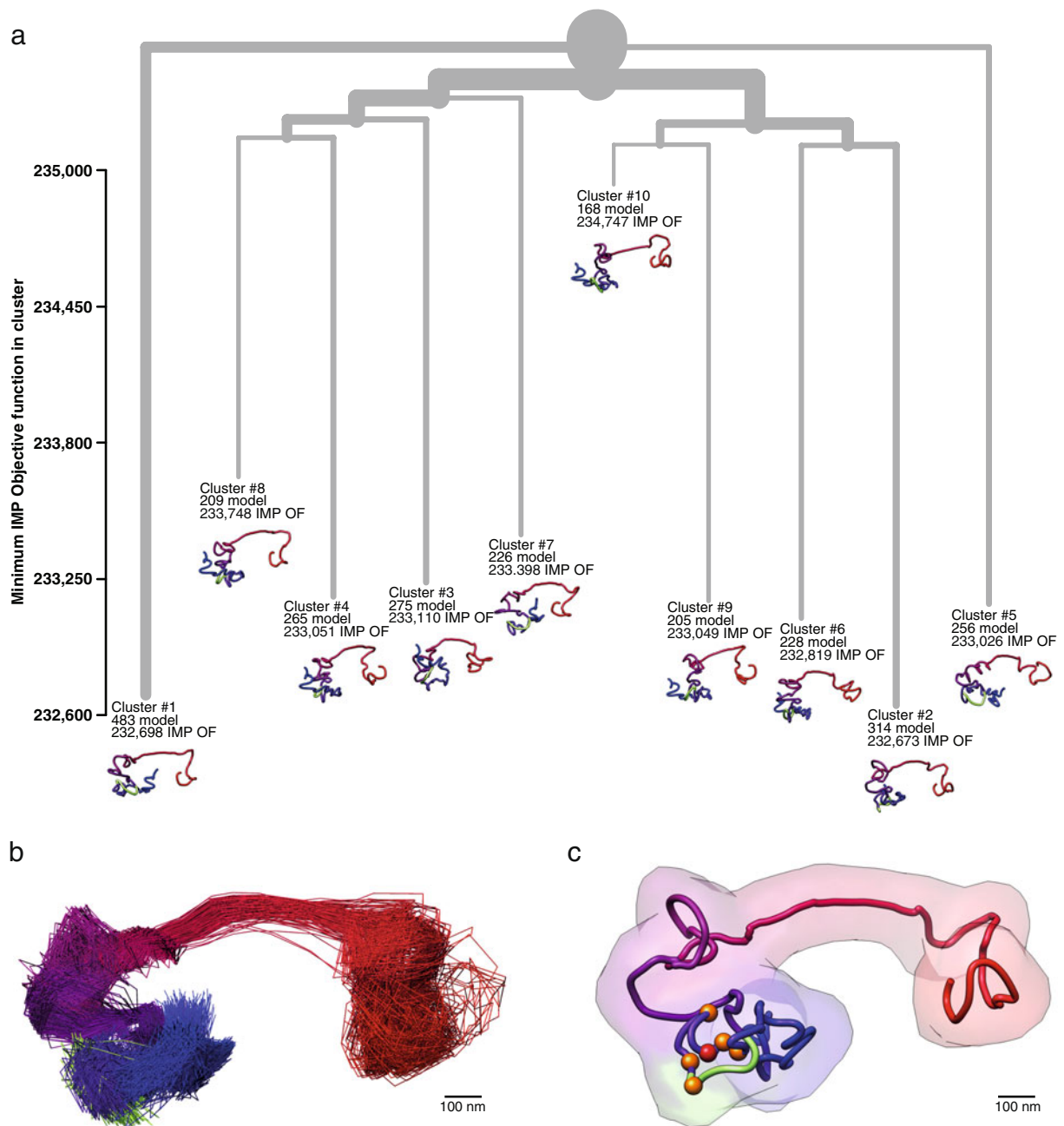




the solutions from the top clusters reflected a range of multiple distinct conformations, which may represent the chromatin conformational differences in a cell population. A density map that covered 80% of the particles in cluster number 2, which was calculated as a Gaussian function of variable standard deviation, indicated that the effective resolution of the  $\alpha$ -globin models was 175 nm (Fig. 2c).

### 3D models of the $\alpha$ -globin domain reveal the formation of chromatin globules

Our 3D models, which were validated by FISH experiments (Bau et al. 2010), accurately reflected the known long-range interactions between the  $\alpha$ -globin genes and their distant regulatory elements. The 3D structure of the  $\alpha$ -globin domain showed the presence of a higher-order chromatin folding motif in



**Fig. 2** Ensemble solutions for the  $\alpha$ -globin 3D models. **a** Cluster analysis for the selected 10,000 models showing the structural relationship between the top ten cluster centroids. The tree was generated based on the structural similarity between each of the centroids. The branch thickness is proportional to the number of solutions at each branch point. Each centroid, colored as in its linear representation (Fig. 1a), is vertically

placed proportional to the lowest IMP objective function within the cluster it represents. **b** Ensemble of solutions in cluster number 2. **c** Model representation for the centroid of cluster 2, which is colored as in its linear representation (Fig. 1a) surrounded by a transparent surface calculated with a Gaussian of 175 nm resolution. Images of the structures shown were generated using the Chimera program (Pettersen et al. 2004)

which groups of adjacent genes clustered together to form what we termed “chromatin globules.” The general features of such chromatin globules included the existence of a limited number of chromatin loops of 50–70 Kb with an average path length of ~500 nm, and anchoring points separated by ~100–200 nm. Analysis of the internal architecture of these globules revealed that active genes were enriched in the cores of these structures. These observations suggested that chromatin globules could represent subnuclear structures dedicated to gene expression, perhaps related to the clustering of shared transcription machineries.

### Future directions

We have shown here that the probabilistic localization of DNA loci in the interphase nucleus can be mapped onto 3D models by satisfying spatial restraints derived from 5C interaction frequency matrices. Although at present, restraints used in IMP were mainly derived from 5C interaction data and chromatin physical properties from the literature, the IMP's

conceptual framework allows the integration of data from different experimental sources (as an example of such techniques see Table 1); however, such observations need to be adequately represented. That is, a trade-off between the level (i.e., the detail) of the representation of an observation in the system and the computational time has to be set in order to allow the exploration of the conformational space. We aim at further developing IMP to accept spatial restraints derived from a variety of experiments, which may result in higher resolution models. Data from FISH and high precision epifluorescence microscopy, for example, could provide a sample of the spatial distance distribution between loci within domains under investigation. These data would ease the IMP calibration necessary for mapping 5C frequencies onto Euclidean distances. Finally, to further improve the resolution and accuracy of our models, a proper treatment of chromatin polymer physics in agreement with 3C-based data (Rosa et al. 2010) will be required within IMP. Current methods, such as the worm-like chain model, among others (Langowski and Heermann 2007), have proven valuable for understanding general features of chromatin fibers, including flexi-

**Table 1** Experimental observations on the structure of genomic domains

3D		1D
Loci to genomic domains	Chromosomes to entire genome	Genome-wide
3C	FISH	ChIP-chip -seq
4C	Tomography	DamID
5C	X-Ray Microscopy	DNaseI-chip -seq
Hi-C		Transcription profiling
FISH		Replication profiling

The figure shows three horizontal scales. The top scale is 'DNA length' in 'nt' (nucleotides), ranging from  $10^0$  to  $10^9$ . The middle scale is 'Volume' in  $\mu\text{m}^3$ , ranging from  $10^{-9}$  to  $10^3$ . The bottom scale is 'Resolution' in  $\mu\text{m}$ , ranging from  $10^{-3}$  to  $10^{-1}$ .

Length, volume and resolution scales adapted from (Langowski and Heermann, 2007). Refer to (Naumova and Dekker, 2010) for a recent review on experimental 3D and 1D technologies.

bility, compaction, and unpacking (Rosa and Everaers 2008).

**Acknowledgments** We thank the Dekker group for their support during the development of our approach. We also thank the IMP community (especially the Sali Lab) and the Chimera developers (<http://www.cgl.ucsf.edu/chimera>). Financial support from the Spanish Ministerio de Ciencia e Innovación (BIO2007/66670 and BFU2010/19310) is also acknowledged. This review was partially based on the authors' previous work (Baù et al. 2010).

## References

- Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprpto A, Karni-Schmidt O, Williams R, Chait BT, Rout MP, Sali A (2007a) Determining the architectures of macromolecular assemblies. *Nature* 450:683–694
- Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprpto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP (2007b) The molecular architecture of the nuclear pore complex. *Nature* 450:695–701
- Alber F, Forster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* 77:443–477
- Baù D, Sanyal A, Lajoie B, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA (2010) The three-dimensional folding of the  $\alpha$ -globin gene domain reveals formation of chromatin globules. *Nat Struct Biol*. doi:10.1038/nsmb.1936
- Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Kamani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816
- 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 USA* 101:16495–16500
- Dekker J (2006) The three 'C's of chromosome conformation capture: controls, controls, controls. *Nat Methods* 3:17–21
- Dekker J (2008) Mapping in vivo chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction. *J Biol Chem* 283:34532–34540
- Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295:1306–1311
- Dostie J, Dekker J (2007) Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc* 2:988–1002
- Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16:1299–1309
- Dostie J, Zhan Y, Dekker J (2007) Chromosome conformation capture carbon copy technology. *Curr Protoc Mol Biol*, Chapter 21, Unit 21.14
- Duan Z, Andronescu M, Schutz K, Mcilwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS (2010) A three-dimensional model of the yeast genome. *Nature* 465:363
- Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584
- Fraser J, Rousseau M, Shenker S, Ferraiuolo MA, hayashizaki Y, Blanchette M, Dostie J (2009) Chromatin conformation signatures of cellular differentiation. *Genome Biol* 10:R37
- Fritsch C, Langowski J (2010) Chromosome dynamics, molecular crowding, and diffusion in the interphase cell nucleus: a Monte Carlo lattice simulation study. *Chromosome Res*. doi:10.1007/s10577-010-9168-1
- Gerchman SE, Ramakrishnan V (1987) Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. *Proc Natl Acad Sci USA* 84:7802–7806
- Higgs DR, Wood WG (2008) Long-range regulation of alpha globin gene expression during erythropoiesis. *Curr Opin Hematol* 15:176–183
- Higgs DR, Vernimmen D, Hughes J, Gibbons R (2007) Using genomics to study how chromatin influences gene expression. *Annu Rev Genomics Hum Genet* 8:299–325
- Hughes JR, Cheng JF, Ventress N, Prabhakar S, Clark K, Anguita E, De Gobbi M, De Jong P, Rubin E, Higgs DR (2005) Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc Natl Acad Sci U S A* 102:9830–9835
- Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
- Lajoie BR, Van Berkum NL, Sanyal A, Dekker J (2009) My5C: web tools for chromosome conformation capture studies. *Nat Methods* 6:690–691
- Langowski J, Heermann DW (2007) Computational modeling of the chromatin fiber. *Semin Cell Dev Biol* 18:659–667
- 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 (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–293
- Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R (2006) Interchromosomal interactions and olfactory receptor choice. *Cell* 126:403–413
- Mateos-Langerak J, Bohn M, De Leeuw W, Giromus O, Manders EM, Verschure PJ, Indemans MH, Gieman HJ, Heermann DW, Van Driel R, Goetze S (2009) Spatially confined folding of chromatin in the interphase nucleus. *Proc Natl Acad Sci U S A* 106:3812–3817
- Miele A, Dekker J (2009) Mapping cis- and trans- chromatin interaction networks using chromosome conformation capture (3C). *Methods Mol Biol* 464:105–121
- Mirny L (2010) The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* doi:[10.1007/s10577-010-9177-0](https://doi.org/10.1007/s10577-010-9177-0)
- Naumova N, Dekker J (2010) Integrating one-dimensional and three-dimensional maps of genomes. *J Cell Sci* 123:1979–1988
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612
- Rosa A, Everaers R (2008) Structure and dynamics of interphase chromosomes. *PLoS Comput Biol* 4:e1000153
- Rosa A, Becker NB, Everaers R (2010) Looping probabilities in model interphase chromosomes. *Biophys J* 98:2410–2419
- Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, De Wit E, Van Steensel B, De Laat W (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38:1348–1354
- Takizawa T, Meaburn KJ, Misteli T (2008) The meaning of gene positioning. *Cell* 135:9–13
- Van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* (39). pii:1869
- Wachsmuth M, Caudron-Herger M, Rippe K (2008) Genome organization: balancing stability and plasticity. *Biochim Biophys Acta* 1783(11):2061–2079
- Wagner G, Braun W, Havel TF, Schaumann T, Go N, Wuthrich K (1987) Protein structures in solution by nuclear magnetic resonance and distance geometry. The polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. *J Mol Biol* 196:611–639
- Wedemann G, Langowski J (2002) Computer simulation of the 30-nanometer chromatin fiber. *Biophys J* 82:2847–2859
- Wurtele H, Chartrand P (2006) Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended chromosome conformation capture methodology. *Chromosome Res* 14:477–495
- Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38:1341–1347