Biological Background

The objects used in our modeling approach are based on well-known biological structures. For low resolution modeling we use topologically associating domains (TADs) for Hi-C data and chromosome contact domains (CCDs) for ChIA-PET data. TADs are megabase-sized regions with abundant interactions throughout the region and substantially fewer interactions with other regions, which appear as clearly discernable blocks along the diagonal of a Hi-C heatmap. CCDs are megabase-sized regions of dense ChIA-PET interactions, with no more than a few interactions between CCDs. For high resolution modeling we use the anchor positions of chromatin loops. For ChIA-PET we use the anchor positions of the identified PET interactions. Looping interactions are harder to identify from Hi-C data because it is lower resolution, but there are algorithms which attempt to identify these interactions from high resolution (1 kb) Hi-C data.

Experimental Input Data

A ChIA-PET experiment identifies genome-wide interactions mediated by a particular protein of interest, and the main experimental result is the identification of the interaction regions (the 'anchors') and the frequency of interaction between each anchor. The data is then divided into two categories by using a threshold on the interaction frequency - interactions with PET counts below the cutoff are called "singletons" (because the original experiments used a cutoff of two, hence, these were reads with a PET count of one), and interactions with counts above the cutoff are simply called PET interactions. The appropriate cutoff primarily depends on the sequencing depth, and will be different for different libraries. Additional statistical tests may be applied to classify some statistically insignificant PET interactions as singletons, but we do not use this approach with our data because the cutoff alone suffices to identify nearly all significant interactions.

Initially ChIA-PET experiments treated the singletons as noise and discarded them from further analysis, but we have recently shown that singleton data recapitulates the low resolution heatmaps obtained in Hi-C experiments. Thus, we now regard singleton data as representative of large-scale genome organization, and we use singletons to model low resolution genome structures. To do so, we first create a Hi-C-like heatmap by binning the singletons using CCDs as the bins. These heatmaps then give interaction frequencies between CCDs which are used for modeling. (Details of the modeling are given below.)

Hi-C data is usually presented as heatmaps, which we coarse-grain according to TAD locations before using it for low resolution modeling. Deeply sequenced Hi-C libraries can be used to perform high resolution modeling, provided the interactions are classified into singletons and true interactions. There is no universally accepted method for doing so, but we have found it useful to re-analyze raw Hi-C data as if it were ChIA-PET data to identify binding sites and PET counts and perform the classification.

Genome Representation

We take advantage of the multiscale nature of ChIA-PET data by representing the genome as a hierarchical tree structure, where the root node represents the whole genome and subsequent levels represent the genome at increasing resolution - one level corresponds to individual chromosomes, the next level corresponds to individual CCDs or TADs, the next corresponds to PET interaction anchors, and the final level includes 'sub-anchors' which are used to fill in the gaps between the anchors. At each level the chromosome is represented by the classical "beads-on-a-string" model, where the beads correspond to the fundamental biological unit at that level, either whole chromosomes, CCDs (TADs), or anchors. At the sub-anchor level the beads are simply used to fill in the missing sections of the genome and do not have an associated functional entity. We generally use 5-7 evenly spaced beads to define the sub-anchors. The beads at each level have a

parent-child relationship, where each low resolution bead is associated with higher resolution child beads which span the same genomic region.

Heatmap Construction and Normalization

Interchromosomal heatmaps are generated by binning both interchromosomal PET interactions and singletons, using the CCDs or TADs as the bins. Intrachromosomal heatmaps are generated by binning singleton data. It is worth noting that these are unequal bins whose sizes are determined by the sizes of the CCDs. We have argued (see Genome Research paper) that such binning is advantageous because it eliminates biases from splitting single CCDs into multiple bins or assigning differing numbers of bins to different CCDs.

Aftering binning, the heatmaps are first normalized to account for unequal bin size. Given two bins with genome sizes s_i and s_j (in Mb), we define a normalized interaction frequency as $\widehat{f} = f_{ij}/(s_i s_j)$, where f_{ij} is the raw interaction frequency. Next we rescale the heatmap so that each row contains the same total interaction count. Such normalization is based on the idea that each genomic region should have equal "visibility," and is a common normalization technique for Hi-C data.

Structure Reconstruction

The general procedure of structure reconstruction is similar for each level, although level-specific details are provided below. At each level we use the experimental data to define a preferred distance, d_{ij} between each pair of beads, and then define an 'energy' which is a function of these distances and the actual distances between each bead, r_{ij}, namely

$$E(\{r_i\}) = \alpha E_{polymer}(\{r_i\}) + \beta E_{data}(\{r_{ij}\}, \{d_{ij}\}),$$

where the first term includes standard polymer interactions such as stretching and bending energies, and the second term includes all additional interactions imposed by the experimental data. The exact energy function and the method to compute preferred distances is different for each level, as detailed below. We work in a top-down approach, first generating low resolution structures and then using these structures to initialize and constrain higher resolution levels.

Chromosome and CCD Levels

Preferred interaction distances are derived from the heatmap interaction frequencies according to $d_{ij} = c f_{ij}^{-\alpha}$, where f_{ij} is the normalized interaction frequency, c is a scaling constant, and α is the scaling exponent. The scaling constant and exponent are adjustable parameters which need to be tuned for each data set. The scaling constant differs according to features of the heatmap, including bin size and average interaction frequency. The scaling exponent can be approximated by various polymer models, but it is known that the genomes of different species have different scaling exponents. We found $\alpha = 0.5$ works well for our GM12878 dataset, but other values are common in the literature. Some of the calculated preferred distances are unrealistically large for bins with lower interaction frequencies, and we cut off large distances according to $\widehat{d} = min(\xi d_{ij}, d_{av})$, where d_{av} is the average heatmap value and ξ is an adjustable parameter which we generally set between 2 and 3.

The energy function is a simple harmonic potential, $E = \sum_{ij}(r_{ij} - d_{ij})^2$, whereby the constraints are modeled as springs connecting each pair of interacting beads. This function is minimized using a standard Monte Carlo simulated annealing method.

At the chromosome level the beads are initialized randomly in the nuclear space. Simulated annealing is performed for several initial configurations and the configuration with the lowest energy is selected as the best

structure. At the CCD level the beads are initiated at random positions within a sphere of radius R_c with center at the location of the chromosome bead. This initial configuration ensures that CCD beads begin near their preferred position and thus speeds convergence of the simulated annealing algorithm.
We note that we do not currently use inter-chromosome singletons during the CCD level simulation. In principle using these singletons would give structures which more reliably represent the relationship between CCDs in different chromosomes. However, in practice the inter-chromosome data is very noisy, and it is not evident how to select specific haploid interactions from the diploid heatmaps. Thus, the CCD level structures are constructed independently of each other but with proper global positioning.

Anchor Level
Preferred distances are calculated from the PET interaction frequency between anchors as $d_{ij} = \delta + e^{-\beta(f_{ij}+\gamma)}$, where β, γ, and δ are adjustable parameters. This relationship is purely phenomenological and was selected to exponentially weight the PET interactions.
The energy function is identical to that used at the chromosome and CCD levels. Anchor positions are initialized by positioning them randomly in a sphere centered on the CCD. It is worth noting that, by definition, anchors only interact with other anchors in the same CCD. This allows the configuration of each anchor group to be determined independently, which greatly decreases the optimization time.

Sub-anchor Level
At the sub-anchor level we consider several contributions to the energy in order to properly generate the loops between anchors. First, in order to ensure that the physical size of a loop scales with its genomic span, we include a term which imposes a larger distance between sequential sub-anchor beads separated by a larger genomic span. All polymer models predict a power law relationship between arc length and physical size, and thus we use $d_{i,i+1} = N^{\alpha}_{i,i+1}$, where $N_{i,i+1}$ is the number of base pairs between sub-anchors i and i+1. These preferred distances contribute a term $E_{dist} = \sum(r_{i,i+1} - d_{i,i+1})^2$ to the total energy. Next, we include a bending energy, which prevents excessive curvature. This energy is $E_{bend} = \frac{1}{2}\sum(1 - \hat{v}_{i,i-1} \cdot \hat{v}_{i,i+1})$, where $\hat{v}_{i,i+1}$ is the unit vector pointing from sub-anchor i to sub-anchor i+1.
It is known that chromatin interaction loops are heavily influenced by the orientation of CTCF binding motifs (Rao et al. 2014; Tang et al. 2015). Based on this observation we reasoned that such sequence specific orientation in chromatin looping could dictate the forms of loops, and thus propose the “hairpin” loops for convergent motifs and the “coiled” loops for tandem motifs, respectively. We introduced the orientation of CTCF binding motifs into our computational algorithm in the following way. The genomic orientation of a motif (determined by whether the motif is directed upstream or downstream) is reflected in the structural, 3D motif orientation, which is defined as a unit vector tangent to the chromatin curve at the location of the motif. The vector points either “along” the fiber (from the 5’ to 3’ direction) or in the opposite direction, depending on the genomic motif orientation. These directions correspond to rightward and leftward motifs, respectively. We assume a pair of interacting anchors with CTCF motifs will preferentially align with their tangent vectors pointing in the same spatial direction. To account for this interaction we include a third energy term based on the orientation of interacting anchors, $E_{orn} = \sum_{i,j \in P}(1 - \hat{\sigma}_i \cdot \hat{\sigma}_j)$, where $\hat{\sigma}_i$ is the orientation of the anchor i and P is a set of pairs of interactions in the current CCD.
These energy terms can be used to model smooth, circular loops passing through the fixed anchor beads, but they do not account for interactions between sub-anchors in different loops. To determine the effect of these

interactions we build two heatmaps. First we use the intra-CCD singletons to construct a sub-anchor heatmap. This heatmap is not directly used to compute preferred distances because it contains many null entries, which are simply consequences of the sparseness of interaction data at extremely high resolutions. To impute these missing values we construct several structures using just the distance and bending energies. For each structure we construct a heatmap using the distance between each pair of loci, and then these heatmaps are averaged to produce a consensus distance heatmap. Each entry in the distance map is then decreased in proportion to the corresponding entry in the singleton heatmap to generate a refined distance heatmap, and these reduced distances are used to define the fourth energy term, $E_{heat} = \sum(r_{ij} - d_{ij})^2$. This term is similar to E_{dist} but the sum is over all pairs of beads instead of just over neighbors.

Combining these terms we arrive at $E_{subanchor} = w_{dist}E_{dist} + w_{bend}E_{bend} + w_{orn}E_{orn} + w_{heat}E_{heat}$, where w_{dist}, w_{bend}, w_{orn}, and w_{heat}, are weights assigned to particular energy terms.

Optimization algorithm

Monte Carlo simulated annealing proceeds in the conventional fashion, namely, at each step a random bead is chosen and shifted by a vector drawn at random from a sphere of a specified radius. The new energy is calculated and the move is accepted if $E_{new} < E_{old}$. If $E_{new} > E_{old}$, then the move is accepted with probability $P = exp(-E_{new}/(T\ E_{old}))$, where T is analogous to the temperature. This form differs from the Boltzmann form typically used in Metropolis Monte Carlo simulations, but any form is acceptable for simulated annealing and this form is convenient because it is insensitive to the magnitude of the energies and thus provides more flexibility with parameter choices. The “temperature” is initialized to $T_{init} > 0$, and is reduced after each step, $T_{new} = \kappa\ T_{old}$, for some $\kappa < 1$. The simulation is checked every $N_{milestone}$ steps, and the simulation is stopped when the energy decrease since the last milestone is below a user defined threshold.