Supplementary Information

1. Supplementary Data. Technical details of PGS.

Supplementary Information

Technical details of PGS

Bin Level Probability Matrix

The released PGS package has been modified slightly from the version used originally in our previous work[1]. One modification is to improve the speed and model resolution. Migrating from older to a new version of IMP (https://integrativemodeling.org/) has increased the speed at least by two fold. Therefore, we are able to increase the resolution accordingly. In the example test data, we provide TAD-level modeling starting from a 100 kb resolution Hi-C matrix. One option of the software requires an input of a raw Hi-C contact frequency matrix. With this option, PGS needs to process the raw matrix such as removing outliers (described previously[1]) and performing normalization (with KR-normalization[2]). We first convert the contact frequency to a contact probability between domain pairs. By definition chromatin regions within TADs show higher interaction frequencies than contacts between chromatin regions between TADs. There are cases in TAD-resolution contact frequency matrix that very loose interaction patterns between neighboring TADs can occur, which suggests a low chance for those consecutive genomic regions to form close contact in 3D space. In contrast to our previous approach[1], consecutive TADs in our current model do not necessarily form contacts between them in 100% of structures in the population. Therefore we now adapt a different strategy for the parameter f^{max}, (i.e. the contact frequency value at which two domains have a 100% probability to form a contact). It serves also as a simple normalization factor that transforms a contact frequency matrix into a contact probability matrix, which then can be used for input in our 3D modeling method. In our previous approach, the *fmax* parameter was unique for each bin and determined by the direct neighbor contacts. In the current method, *fmax* is a uniform scaling constant. A bin level contact probability matrix, denoted as $\boldsymbol{P} = (p_{ij})_{K\times K}$, can be calculated through the formula $p_{ij} = \min\left(\frac{f_{ij}}{f^{max}}, 1\right)$ describing the probability of contact between region *i* and *j*, where f_{ij} and p_{ij} represents their contact frequency and probability values, respectively.

The choice of f^{max} will affect the scale of global contact frequencies, and it depends on the data set. Although we think that choosing the right f^{max} will result in consistent observed contact frequency observed between model and other non-Hi-C-based experiments, the relative contact frequencies between different TAD-TAD pairs will mostly not be affected by tuning the f^{max}. Our experience show that at saturation (where no more contact restraints can be satisfied), a TAD is surrounded by ~21-25 other TADs. The value of f^{max} is then chosen so that the average contact probability sum of a TAD is about 23. From our experience, such value of fmax will lead to low restraints violation in the structure optimization down to a_{ij} ~ 1% and the number of contact restraints has reach saturation (non-tolerable violation score if more restraints are added).

TAD Level Probability Matrix

As described in our previous work[1],a contact between two domains is defined by the contact frequencies of the (bin level) chromatin segments between both domains. We define TAD level contact probability $\boldsymbol{A} = (a_{ij})_{N\times N}$, where a_{ij} is the contact probability between TAD *i* and *j*, and N is the total number of TADs.

If we define mapping $b(i)$ is the set of all bins in matrix ***P*** that belong to TAD *i*, we can calculate matrix ***A*** by

$$a_{ij} = mean\langle top10\%\{p_{\alpha\beta}|\alpha \in b(i), \beta \in b(j)\}\rangle$$

Here discarded bins such as centromeres are excluded from the calculation. In addition, normalization will sometimes cause blowouts that some contacts are extremely higher than surrounding contacts. These contacts are identified as outliers by $\{p \mid p > \mu + 1.5IQR \; or \; p < \mu + 1.5IQR\}$, where $p \in \{p_{\alpha\beta}|\alpha \in b(i), \beta \in b(j)\}$ and $\mu = mean\{p_{\alpha\beta}|\alpha \in b(i), \beta \in b(j)\}$, IQR is the interquartile range of $\{p_{\alpha\beta}\}$. Outliers will also be excluded from calculation.

Technical detail about the dynamics process

Modifications in the dynamic simulation technique of the M-step. PGS now uses genome structure coordinates from a previous iteration step as starting configurations to reduce the search space of local optima in the next M-step. To make the optimization more efficient, at initial optimization steps the nuclear volume is first expanded and then gradually shrunk to its normal value while performing simulated annealing dynamics (e.g. setting a nuclear radius (Rnuc) from 1.2 to 0.8 Rnuc with interval of 0.1 Rnuc). Our experience shows that this strategy helps to reach an optimum conformation more quickly.

The lack of constraints at the very earliest A/M steps usually causes extended conformations of chromosomes. To handle this problem, we introduced a bounding spherical volume for every chromosome to mimic chromosome territory applied only at the very first stage of the A/M optimization. The radius of the bounding sphere is proportional to the chromosome length. This spherical territory constraint is only applied at the very early stage of A/M optimization and is not applied at later stages of the optimization. This strategy helps both homologues copies to have similar distribution of contact constraints during the optimization.

References

1. Tjong, H. et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl. Acad. Sci. USA* **113**, E1663-72 (2016).

2. Knight, P.A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J Numer Anal* **33**, 1029-1047 (2013).