**Supplemental document**

**Contents**

[1.1 Preparation of synthetic datasets from a yeast genome for validation 1](#_Toc435434832)

[1.2 Validation on the synthetic chromosomal contact data of the yeast genome 2](#_Toc435434833)

[1.3 Data normalization 3](#_Toc435434834)

[1.4 Description and Configuration of Parameters 5](#_Toc435434835)

## Preparation of synthetic datasets from a yeast genome for validation

In addition to testing our method on the independent synthetic dataset published in (Trussart et al., 2015) and on the real Hi-C data of the human B-cells (Lieberman-Aiden et al., 2009a), we used the yeast genome structure published in (Duan et al., 2010), excluding the chromosome 12 whose structure could not be observed in practice, to generate synthetic chromosomal contact matrices to validate our method. The whole genome structure is represented by 4,342 points (fragments) at 50KB resolution. IFs between the fragments were obtained using the formula , where is the Euclidian distance between fragment in the structure, and is the average of all distances. The role of is to avoid numerical issues when IFs are too small and is set to 2. It is expected that reconstructed structures from these IFs are scaled versions of the true structure.

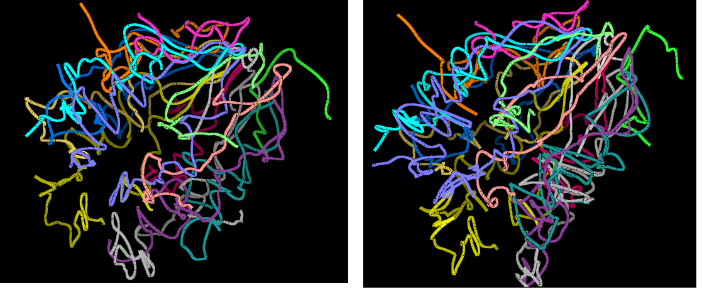
Furthermore, in order to test the robustness of MOGEN, noise was introduced into the contact matrix to generate additional datasets in the following manner: some percent of pairs of fragments between 0.2 % and 30% were randomly selected without replacement, and for each selected pair of fragments, their IF values ( and , ) were adjusted as follows, and , where is a random number between . This adjustment of IF values not only can introduce noise to the IF values of contacts, but also can make contacts into non-contacts if an IF value is reduced to below the threshold or vice versa. And without any normalization and pre-processing, these noisy contact matrices derived from the known genome structure were directly used to generate genome structural models.

To avoid the problem of scaling in superimposing large structures when comparing reconstructed structures to the true structure, similar to (Lesne et al., 2014), we used the Pearson’s correlation between reconstructed distances and true distances (i.e., the distance correlation) to evaluate the accuracy of the reconstructed structural models.

## Validation on the synthetic chromosomal contact data of the yeast genome

The synthetic dataset in (Trussart et al., 2015) contains only one chromosome, so we also tested our tool on synthetic datasets of a genome consisting of several chromosomes. The preparation of these synthetic datasets including how the noise was produced are described in 1.1. Since MOGEN may generate different structures in different runs due to the random initialization, for each dataset, we generated an ensemble of 50 structures for analysis. The distance correlations between the generated structures in an ensemble and the true structure were calculated and then their average was taken as the final accuracy. When there is no noise, the average distance correlation of structures generated by MOGEN is about 0.96. As the percentage of noise increases to 10%, it decreases just a little (Figure 1), but when the percentage of noise is higher than 10%, the correlation starts to drop to below 0.9. Generally, the results suggest that MOGEN is capable of reconstructing reasonable models from noisy data. Figure 2 shows the true structure (left) and a reconstructed structure (right) with 6% noise. The contact matrices and 3D structures generated by MOGEN can be found at <http://calla.rnet.missouri.edu/mogen/>.

**Figure 1.** Distance correlations between reconstructed models and the true structure.

****

**Figure 2.** True structure (left) and reconstructed structure (right) with 6% noise in the synthetic contact data.

## Data normalization

We employed a simple data normalization protocol (Lieberman-Aiden et al., 2009b) to pre-process the Hi-C data. Given an initial *n x n* interaction frequency matrix C representing contact numbers between n units of a chromosome that were generated from a raw Hi-C dataset, an element denoting IF between regions and of a chromosome in a normalized matrix is calculated according to the formula below.

The first term, in the initial IF matrix, is the observed number of Hi-C reads linking regions and . The inverse of the second term in the formula is the expected IF between two regions and , which is used to normalize . The expected IF is the product of the number of interactions of each region divided by the total number of interactions. The normalized IF () is obtained by dividing the observed IF () by the expected IF. This normalized IF could be interpreted as the likelihood ratio (odds) of regions and in contact, which is the number of observed reads divided by the number of expected reads between the two regions. We called this normalization “coverage normalization” in our implementation for normalization (called normalizer) provided in the software package.

Our normalizer also implements a more sophisticated normalization (Imakaev et al., 2012) that we called “iterative normalization”. This normalization doesn’t assume specific sources of biases, such as in (Cournac et al., 2012a). However, it makes an assumption that the bias between 2 regions can be represented as the product of individual biases from each region. In implementation, this iterative normalization just repeats the above coverage normalization until the matrix is converged. Intra-chromosomal contacts and inter-chromosomal contacts were normalized separately when applying this normalization (Imakaev et al., 2012). We were unsure how to incorporate intra-chromosomal contacts and inter-chromosomal contacts after normalization so that we didn’t use this normalization for the whole genome. In our experience, this normalization is more effective than the coverage normalization when applied for intra-chromosomal contacts (producing higher quality models).

In the future, we plan to implement more normalization methods for our normalizer, such as (Cournac et al., 2012b), (Hu et al., 2012), and perform a comparison between normalization methods.

## Description and Configuration of Parameters

There are three types of parameters, parameters related to the chromosomal contact data, parameters of the objective function and system parameters to configure how the tool is run. These parameters can be specified in a text file, parameters.txt – which is required to run the tool, in any order and in the form of pairs of a key and a value (key = value). This file can contain comments starting with ‘#’, which are ignored by the tool.

The first parameter related to the chromosomal contact data is the number of chromosome in the contact data. For chromosome reconstruction, it is always one. If the actual number of chromosomes in the input file is different from this number, an exception will be thrown. This number is specified by the key, NBR\_OF\_CHR. The highest indices of fragments of chromosomes are needed to identify fragments of chromosomes. These numbers are put in a separate text file with the name of the file specified by the value of the key, CHR\_UPPER\_BOUND\_ID\_FILE, in the parameters.txt file. Each line of this text file contains two numbers, the first number is index of the chromosome (1 - 24) and the second number is the highest fragment index of the corresponding chromosome. Chromosome indices in this file must be consistent with the number of chromosomes specified by the value of the key, NBR\_OF\_CHR.

Because of the noise, and the fact that the input data is generated from an ensemble of millions of cells, almost all fragments (at low resolutions) are in contact, which could be unrealistic for structure of a single cell. Therefore, it is necessary to apply thresholds to chromosomal contacts to remove contacts with low IF (unreliable contacts) before the reconstruction. The method implemented by this tool can tolerate noise contacts well and especially can reconstruct similar structures for a range of threshold of intra-chromosomal contacts. Generally, thresholds for intra-chromosomal and inter-chromosomal contacts should be similar or not too far from each other to avoid bias among inter-and intra-chromosomal contacts. Users are suggested to try different thresholds for both intra- and inter-chromosomal contacts. The keys, INTRA\_IF\_THRESHOLD and INTER\_IF\_THRESHOLD, are used to specify thresholds for intra-chromosomal contacts and inter-chromosomal contacts respectively. Contacts with IF less than the corresponding threshold will be treated as non-contacts.   
Distance thresholds are needed for the objective function in (Trieu and Cheng, 2014) and defined by five parameters. ADJACENT\_DIST, CONTACT\_DIST and POS\_MIN\_DIST define the square maximum distance between two adjacent fragments of the same chromosome, the maximum square distance for fragments that are in contact and the minimum square distance between any two fragments respectively. NEG\_MAX\_DIST\_INTRA and NEG\_MAX\_DIST\_INTER specify the square maximum distance for any two fragments of the same chromosome and the square maximum distance for any two fragments that are not in the same chromosome. Theoretically, NEG\_MAX\_DIST\_INTER should be around the square of the diameter of the nucleus. Ideally, the values of these parameters should be calibrated from FISH measurements of the same cell type. However, the method can tolerate a range of value of the contact distance (CONTACT\_DIST), FISH measurements of cells from the same species could be used to estimate the contact distance and then other distances can be set relatively to the contact distance. In (Trieu and Cheng, 2014) values of these parameters were calibrated from the previous study in (Mateos-Langerak et al., 2009).

Adjustable parameters for the terms of the objective function are defined in separate text files. The key, POS\_MAX\_DIST\_WEIGHT\_FILE, specifies the file containing weight factors for the maximum distance constraints of contacts (the distance between fragments that are in contact should be less than a distance), - in the scoring function. The first line of the file contains a number specifies the weight for the maximum distance constraint of all inter-chromosomal contacts. Each of following lines contains two numbers, the first number is the chromosome index and the second number is the weight for the maximum distance constraint of intra-chromosomal contacts of the chromosome. Higher weights will give higher priority for contacts to be satisfied. However, the number of satisfied non-contacts will decrease. In our work, these weights are all set at value of 1 and weights of other constraints will be adjusted. The key, POS\_MIN\_DIST\_WEIGHT\_FILE, specifies the file containing weight factors for the minimum distance constraints of contacts (the distance between fragments that are in contact should be larger than a distance), - in the scoring function. Similarly, the first line of the file contains a number for the weight factor of the minimum distance constraint of inter-chromosomal contacts and each of following lines contains two numbers, the first number is the chromosome index and the second number is the weight for the minimum distance constraint of intra-chromosomal contacts of the chromosome. These weights prevent fragments from being too close to each other and generally don’t affect to the scores of structures. In our work, these weights have the same value. To adjust these weights, initially, they can be set at a random value (one is suggested), then, depending on how close fragments are in the 3D structures, the weights can be increased or decreased accordingly.

The key, NEG\_MIN\_DIST\_WEIGHT\_FILE, is to specify the file with the same format as the two previous file and containing weight factors for the minimum distance constraints of non-contacts (the distance between fragments that are not in contact should be larger than a distance), - in the scoring function. These weights are to prioritize non-contacts to be satisfied. When non-contacts are satisfied more, the number of satisfied contacts will be decreased. Therefore, these weights and the weights of the maximum distance constraints of contacts are the most frequently adjusted parameter to balance the number of satisfied contacts and the number of satisfied non-contacts. Since, chromosomes are supposed not to intermingle, inter-chromosomal contacts should be rare and therefore it is desired that inter-chromosomal non-contact score of pairs of chromosomes are high. Because of the abundance of inter-chromosomal non-contacts in the input data, these scores can be high easily. On the other hand, we would like to satisfy as many as possible inter-chromosomal contacts, so that the weight of inter-chromosomal non-contacts is usually less than one. Initially, it can be set at one, and then gradually decreased until inter-chromosomal non-contact scores of pairs of chromosomes become too low (less than 50%) – indicating that chromosomes start to intermingle. The weights for chromosomes are adjusted to balance contact and non-contact scores of corresponding chromosomes. Typically, these weights are larger than one, because intra-chromosomal contacts often have high IF and can be satisfied easily, so that these weights are set to increase the number of satisfied non-contacts.

The key, NEG\_MAX\_DIST\_WEIGHT\_FILE, specifies the file with the same format as the three previous files and containing weight factors for the maximum distance constraints of non-contacts (the distance between fragments that are not in contact should be less than a distance), - in the scoring function. These weights are to keep structures to stay within a certain diameter (nucleus diameter for genome structures). Typically, the weights for chromosomes are the same and larger than one. The weight of inter-chromosomal non-contacts is usually less than one to make chromosomes to stay far from each other.

The values of *W*1, *W*2, *W*3 and *W*4 for the normal B cell data (GM06990) used to reconstruct 3D genome structures are reported below.

1. for all pairs of chromosomes .
2. for pairs of chromosome .
3. for pairs of chromosome where and when .
4. for pairs of chromosome where

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|  | 1.2 | 1.35 | 2.2 | 2.2 | 2.3 | 2.6 | 3.8 | 3.9 | 3.3 | 4.8 | 4.2 | 4.0 | 10.0 | 6.0 | 6.0 | 6.5 | 6.5 | 14.5 | 10 | 13.5 | 7 | 9.0 | 3.0 |

Overall, though, the number of parameters that need to be adjusted seems large, in our experiment, only one parameter, the weight of the constraints of inter-chromosomal non-contacts, requires careful adjustment. The other weights can have a wide-range of acceptable values and can be finalized after a few rounds of generating and observing structures. In addition, many of them can have the same value. In case of chromosome reconstruction, only the first line of these four files is read to get corresponding parameters for chromosome reconstruction.

There are seven system parameters. The first parameter is NUM to specify number of structures to be generated. The chromosomal contact data file is specified by the key, INPUT\_FILE. The output folder is defined by OUTPUT\_FOLDER. A prefix for file name of structures is defined by the key, FILE\_PREFIX. Structure files can have a header defined by the key, FILE\_HEADER. To print out information during the reconstruction, the key, VERBOSE, must be set to ‘true’. The initial learning rate can be set by the key, LEARNING\_RATE.

**References**

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., and Mozziconacci, J. (2012a). Normalization of a chromosomal contact map. BMC Genomics *13*, 436.

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., and Mozziconacci, J. (2012b). Normalization of a chromosomal contact map. BMC Genomics *13*, 436.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363–367.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J.S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinforma. Oxf. Engl. *28*, 3131–3133.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

Lesne, A., Riposo, J., Roger, P., Cournac, A., and Mozziconacci, J. (2014). 3D genome reconstruction from chromosomal contacts. Nat. Methods *11*, 1141–1143.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009a). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009b). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Mateos-Langerak, J., Bohn, M., De Leeuw, W., Giromus, O., Manders, E.M.M., Verschure, P.J., Indemans, M.H.G., Gierman, H.J., Heermann, D.W., and Van Driel, R. (2009). Spatially confined folding of chromatin in the interphase nucleus. Proc. Natl. Acad. Sci. *106*, 3812–3817.

Trieu, T., and Cheng, J. (2014). Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. Nucleic Acids Res. gkt1411.

Trussart, M., Serra, F., Baù, D., Junier, I., Serrano, L., and Marti-Renom, M.A. (2015). Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res. gkv221.