Bioinformatics, 33(14), 2017, 2199–2201
doi: 10.1093/bioinformatics/btx152
Advance Access Publication Date: 22 March 2017
Applications Note

Genome analysis

HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps

Koon-Kiu Yan,[1,2,*] Galip Gürkan Yardımcı,[3] Chengfei Yan,[1,2] William S. Noble,[3,4] and Mark Gerstein[1,2,5,*]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA, [3]Department of Genome Sciences, [4]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA and [5]Department of Computer Science, Yale University, New Haven, CT, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on November 16, 2016; revised on March 11, 2017; editorial decision on March 13, 2017; accepted on March 21, 2017

Abstract

Summary: Genome-wide proximity ligation based assays like Hi-C have opened a window to the 3D organization of the genome. In so doing, they present data structures that are different from conventional 1D signal tracks. To exploit the 2D nature of Hi-C contact maps, matrix techniques like spectral analysis are particularly useful. Here, we present HiC-spector, a collection of matrix-related functions for analyzing Hi-C contact maps. In particular, we introduce a novel reproducibility metric for quantifying the similarity between contact maps based on spectral decomposition. The metric successfully separates contact maps mapped from Hi-C data coming from biological replicates, pseudo-replicates and different cell types.

Availability and Implementation: Source code in Julia and Python, and detailed documentation is available at https://github.com/gersteinlab/HiC-spector.

Contact: koonkiu.yan@gmail.com or mark@gersteinlab.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide proximity ligation assays such as Hi-C have emerged as powerful techniques to understand the 3D organization of the genome (Kalhor *et al.*, 2011; Lieberman-Aiden *et al.*, 2009). Although these techniques offer new biological insights, they demand different data structures and present new computational questions (Ay and Noble, 2015; Dekker *et al.*, 2013). For instance, a fundamental question of practical importance is, how can we quantify the similarity between two Hi-C data sets? In particular, given two experimental replicates, how can we determine if the experiments are reproducible?

Data from Hi-C experiments are usually summarized by so-called chromosomal contact maps. By binning the genome into equally sized bins, a contact map is a matrix whose elements store the population-averaged co-location frequencies between pairs of loci. Therefore, mathematical tools like spectral analysis can be extremely useful in understanding these chromosomal contact maps. Our aim is to provide a set of basic analysis tools for handling Hi-C contact maps. In particular, we introduce a simple but novel metric to quantify the reproducibility of the maps using spectral decomposition.

2 Algorithms

We represent a chromosomal contact map by a symmetric and non-negative adjacency matrix W. The matrix elements represent the frequencies of contact between genomic loci. Recent single-cell imaging experiment suggests that the frequency serves as a reasonable proxy of spatial distance (Wang *et al.*, 2016). In principle, the larger the value of W_{ij}, the closer is the distance between loci i and j. The starting point of spectral analysis is the Laplacian matrix L, which is defined as $L = D - W$. Here D is a diagonal matrix in which $D_{ii} = \sum_j W_{ij}$ (the coverage of bin i in the context of Hi-C). The Laplacian matrix further takes a normalized form $\ell = D^{-1/2}\mathrm{L}\,\mathrm{D}^{-1/2}$ (Chung, 1997). It can be verified that 0 is an eigenvalue of ℓ, and the set of eigenvalues of ℓ $(0 \le \lambda_0 \le \lambda_1 \le \ldots \le \lambda_{n-1})$ is referred to as the spectrum of ℓ.

Given two contact maps W^{A} and W^{B}, we propose to quantify their similarity by decomposing their corresponding Laplacian matrices ℓ^{A} and ℓ^{B} respectively and then comparing their eigenvectors. Let $\{\lambda_0^{\mathrm{A}}, \lambda_1^{\mathrm{A}}, \ldots, \lambda_{n-1}^{\mathrm{A}}\}$ and $\{\lambda_0^{\mathrm{B}}, \lambda_1^{\mathrm{B}}, \ldots, \lambda_{n-1}^{\mathrm{B}}\}$ be the spectra of ℓ^{A} and ℓ^{B}, and $\{v_0^{\mathrm{A}}, v_1^{\mathrm{A}}, \ldots, v_{n-1}^{\mathrm{A}}\}$ and $\{v_0^{\mathrm{B}}, v_1^{\mathrm{B}}, \ldots, v_{n-1}^{\mathrm{B}}\}$ be their sets of normalized eigenvectors. A distance metric S_{d} is defined as

$$S_{\mathrm{d}}(A,B) = \sum_{i=0}^{r-1} \parallel v_i^A - v_i^B \parallel . \qquad (1)$$

Here $\parallel\ \parallel$ represents the Euclidean norm. The parameter r is the number of leading eigenvectors picked from ℓ^{A} and ℓ^{B}. In general, S_{d} provides a metric to gauge the similarity between two contact maps. v_i^{A} and v_i^{B} are more correlated if A and B are two biological replicates as compared with the case when they are two different cell lines (see Supplementary Fig. S1).

For the choice of r, like any principal component analysis, the leading eigenvectors are more important than the lower ranked eigenvectors. In fact, we observe that the Euclidean distance between a pair of high-order eigenvectors is the same as the distance between a pair of unit vectors whose components are randomly sampled from a standard normal distribution (see Supplementary Fig. S2). In other words, the high-order eigenvectors are essentially noise terms, whereas the signal is stored in the leading vectors. As a rule of thumb, we found the choice $r = 20$ is good enough for practical purposes. Furthermore, as the distance between a pair of randomly sampled unit vectors presents a reference, we linearly rescale the distance metric into a reproducibility score Q ranges from 0 to 1 (see the Supplementary Material).

We used HiC-spector to calculate the reproducibility scores for more than a hundred pairs of Hi-C contact maps. As shown in Figure 1, the reproducibility scores between pseudo-replicates are greater than the scores for real biological replicates, which are greater than the scores between maps from different cell lines (see the Supplement). It is worthwhile to point out that two contact maps can be compared in terms of features like topologically associating domains (TADs) and loops. It depends strongly on the choices of methods and parameters. Nevertheless, what we refer to, as 'reproducibility' is a direct comparison of the contact maps.

Mathematically there are different ways to compare two matrices. For instance, one could assume all matrix elements are independent and define a distance metric using Spearman correlation. The intuition behind S_{d} is essentially a better way to decompose a contact map. The normalized Laplacian matrix is closely related to a random-walk-process taking place in the underlying graph of W. The leading eigenvector refers to the steady state distribution; the next few eigenvectors correspond to the slower decay modes of the random walk process and capture the densely interacting domains that are highly significant in contact maps. In fact, HiC-spector can better separate biological replicates and non-replicates compared with the correlation coefficient (see Supplementary Fig. S3).

Apart from the reproducibility score, HiC-spector provides a number of matrix algorithms useful for analyzing contact maps. For instance, to perform a widely used normalization procedure for contact maps (Imakaev *et al.*, 2012), we include the Knight-Ruiz algorithm (Knight and Ruiz, 2012), which is a newer and faster algorithm for matrix balancing. Also, we have included the functions for estimating the average contact frequency with respect to the genomic distance, as well as identifying the so-called A/B compartments (Lieberman-Aiden *et al.*, 2009) using the corresponding correlation matrix.

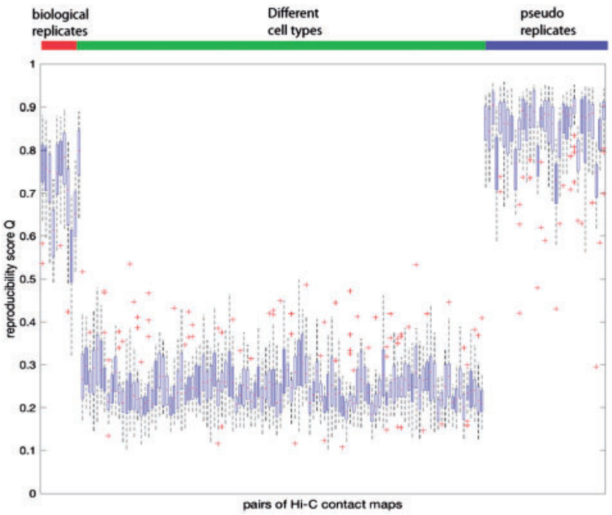

Fig. 1. Reproducibility scores for three sets of Hi-C contact maps pairs. Contact maps came from Hi-C experiments performed in 11 cell lines. Biological replicates refer to a pair of replicates of the same experiment. Pseudo replicates are obtained by pooling the reads from two replicates together performing down sampling. There are 11 biological replicates, 33 pairs of pseudo replicates, and 110 pairs of maps between different cell types. Each box shows for a pair the distribution of Q in 23 chromosomes, with crosses as the outliers

3 Implementation and benchmark

HiC-spector is a library written in Julia, a high-performance language for technical computing. A Python script for the reproducibility score is also provided. The bottleneck for evaluating Q is matrix diagonalization. The runtime is very efficient but depends on the size of contact maps (see Supplementary Fig. S5 for details).

4 Materials and methods

Hi-C data are generated by the ENCODE consortium (see the Supplementary Material). Contact maps were generated using the tool cworld (https://github.com/dekkerlab/cworld-dekker).

Acknowledgements

We want to thank the 3D Nucleome subgroup in the ENCODE consortium for processing the Hi-C data and discussion.

Funding

This work has been supported by NIH award [U41 HG007000].

Conflict of Interest: none declared.

References

Ay,F. and Noble,W.S. (2015) Analysis methods for studying the 3D architecture of the genome. *Genome Biol.*, **16**, 183.

Chung,F. (1997). *Spectral Graph Theory*. American Mathematical Society, Providence, Rhode Island.

Dekker,J. *et al*. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet*., **14**, 390–403.

Imakaev,M. *et al*. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Kalhor,R. *et al*. (2011) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol*., **30**, 90–98.

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

Lieberman-Aiden,E. *et al*. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Wang,S. *et al*. (2016) Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, **353**, 598–602.