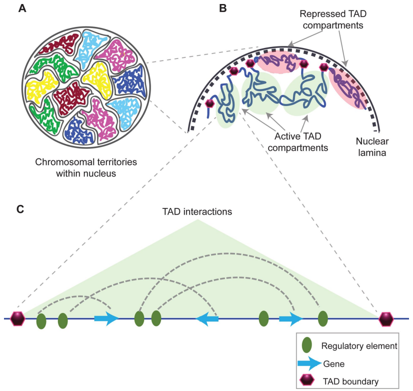
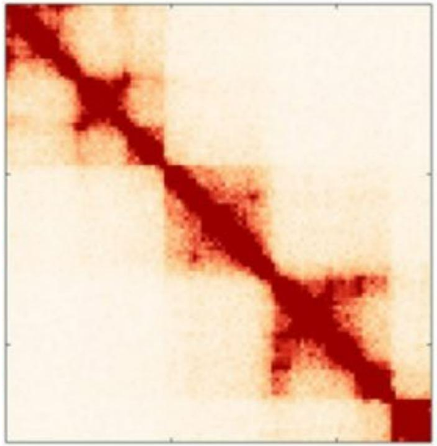
Analysis of TAD Boundary Finding Algorithms Based on Chromosome Hi-C Data

Team Member: Katrina Liu, Linda Zhou, Shreya Varra, Amy Zhu, Leon Xu

Background

- Topological Associated Domains (TADs)
- Chromatin Conformation Capture (3C/Hi-C)





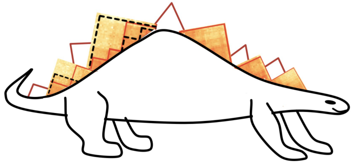
Research Question

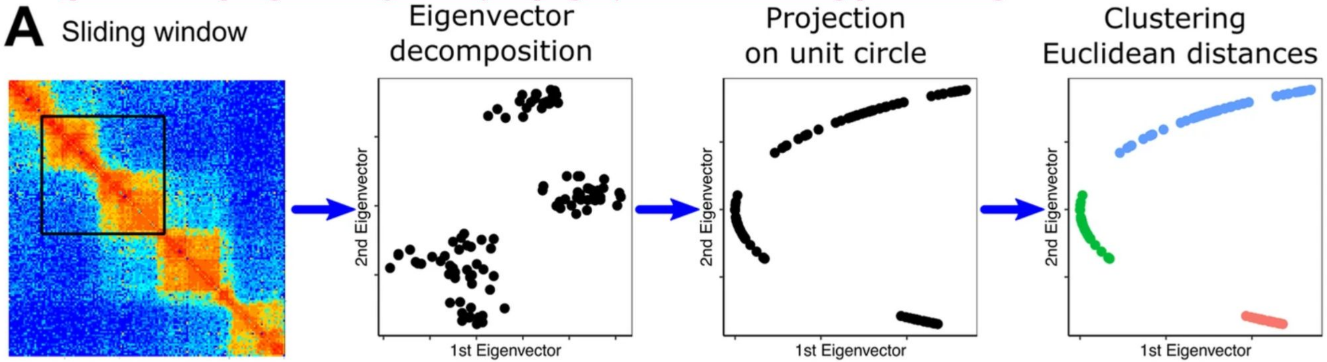
There are several approaches for determining the boundaries of the regions in sample Hi-C data (TAD callers).

We want to analyze and compare the accuracy and efficiency of selected TAD callers in order to determine the best approach for future research needs.

Algorithms for Finding TAD boundaries in Hi-C data

- Armatus (Filippova et al., 2014)
- HiCSeg (Lévy-Leduc et al., 2014)
- SpectralTAD (Cresswell et al., 2020)

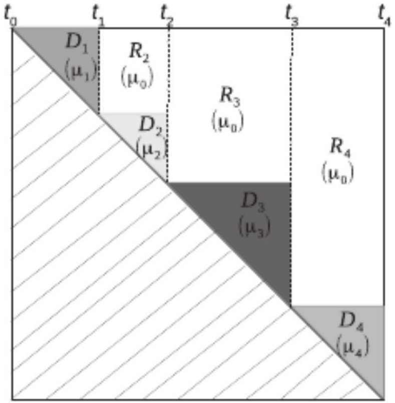




Data Source

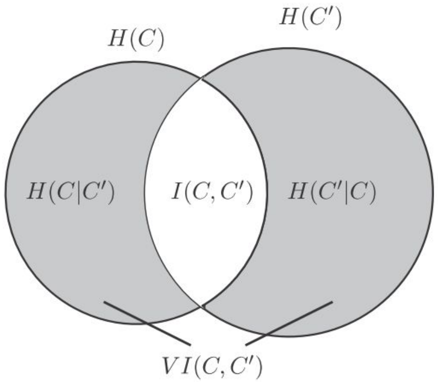
- Synthetic data (Lévy-Leduc et al., 2014)

- Human fibroblast and mouse embryonic data (Dixon et al)



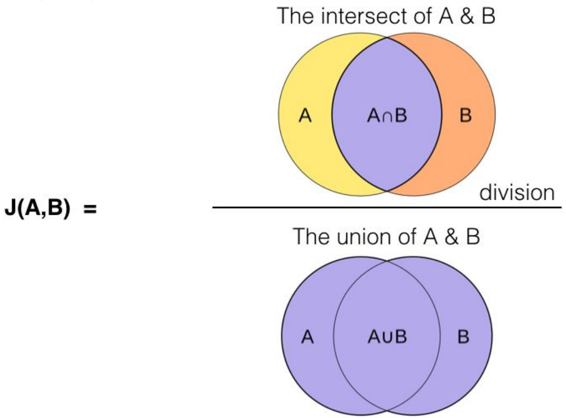
Measuring the Accuracy of Algorithms

- Variation of Information



$$VI(\mathcal{C},\mathcal{C}') = H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C},\mathcal{C}')$$

- Jaccard Index



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Measuring the Efficiency of Algorithms

- Big O
- Timed Execution

Armatus	HiCSeg	SpectralTAD
$O(\lvert D\rvert \log(\lvert D\rvert) + \lvert\Gamma\rvert(n^2 + \lvert D\rvert))$	$O(Kn^2)$	$O(Kw^2nm + w^2n)$

D: the union set of all intervals generated for each resolution
Γ: the set of resolutions
n: the size of Hi-C matrix
m: the number of iterations to convergence for decomposition
w: window size
K: the number of boundary intervals
K (spectralTAD): the number of eigenvectors

Conclusion and Further Discussion

- Spectral TAD (published 2020) > Armatus (2014) ≥ HiCSeg (2014)
- Biological and technological limitations associated with TAD
- Provide a guideline for future references.
- Further impacts

Reference

- Filippova, D., Patro, R., Duggal, G. et al. Identification of alternative topological domains in chromatin. Algorithms Mol Biol 9, 14 (2014). https://doi.org/10.1186/1748-7188-9-14
- Lévy-Leduc, Celine et al. "Two-dimensional segmentation for analyzing Hi-C data." Bioinformatics (Oxford, England) vol. 30,17 (2014): i386-92. doi:10.1093/bioinformatics/btu443
- Cresswell, K.G., Stansfield, J.C. & Dozmorov, M.G. SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. BMC Bioinformatics 21, 319 (2020). https://doi.org/10.1186/s12859-020-03652-w
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012, 485 (7398): 376-80.