Hypergraph Geometry for Multi-Way Chromatin Contacts

Tuning the Inter-Chromosomal Signal in Mouse Embryonic Stem-Cell Pore-C Data

Shivansh Sharma • University of Michigan, Department of Mathematics REU

Project advisors: Prof. Bloch, Joe Huang, Joseph Duhamel

Abstract

Hi-C records chromatin contacts pairwise, so multi-way contacts, in which three, four, or more loci meet simultaneously, escape it. Pore-C long reads capture such contacts directly, and they are the subject of this report. Because a Pore-C read joins several loci at once, the data is naturally a hypergraph, with loci as nodes and each read as a hyperedge. We study a mouse embryonic stem-cell dataset of 195 loci and 29,588 reads, and compute two geometric quantities on the hypergraph: a nonlinear (LogExp) eigenvector centrality and a Forman–Ricci curvature on the containment poset. The curvature partitions the reads into two classes of nearly opposite chromosomal composition, an **interior** class that is about 98% inter-chromosomal and a **bridge** class that is about 85% intra-chromosomal. Because the two are so unequal, mixing them in chosen proportions ρ and κ sets the intra-to-inter ratio to any intermediate value; a 3:1 target, for example, requires only about 5% of the bridge pool. The resulting chromosome-pair graph is dominated by a chromosome-8 hub that persists across every mixing ratio examined. We present the method in full, a worked example that can be verified by hand, and a candid account of what the results do and do not establish.

1. Introduction

A genome is written as a line of bases, but it does not function as one. Roughly two metres of DNA fold into a nucleus a few microns across, and in that folding a locus's position along the sequence ceases to predict its position in space. The two coordinates decouple, and it is the spatial one that governs function. Folding at every scale, from nucleosomes and fibres to loops and whole-chromosome territories, continually brings sequence-distant loci into contact.

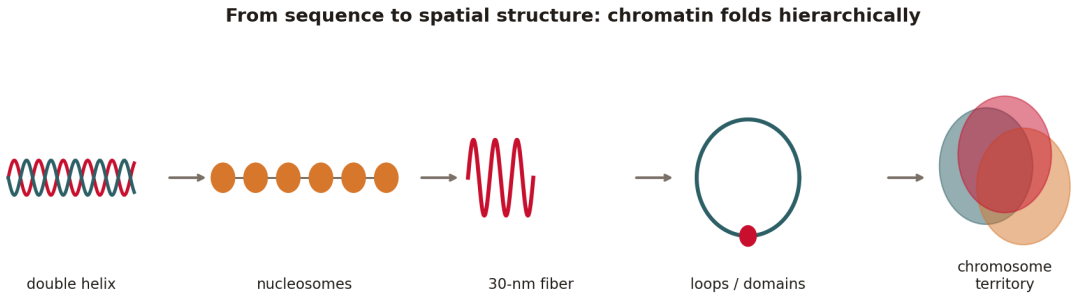


Figure 1: *Chromatin folds hierarchically: from the double helix, to nucleosomes, to the 30-nm fibre, to loops and domains, to chromosome territories. Folding is what brings sequence-distant loci into physical contact.*

Each chromosome occupies its own territory, and the contacts *between* territories describe how those territories are arranged. This inter-chromosomal geometry is the object of interest. It is also the harder signal to recover, since most contacts remain within a single chromosome and dominate the data.

Hi-C, the standard assay, records contacts as pairs. Anything genuinely higher-order, a set of loci in mutual proximity at a single moment, is flattened into a collection of pairwise edges, and the joint structure is lost. Pore-C avoids this: a single long read traverses several co-located fragments and records them together, so one read constitutes one multi-way contact.

This is precisely the setting a hypergraph is built for. An ordinary graph edge has two endpoints, whereas a hyperedge may have any number. We therefore take the loci as nodes and each read as a hyperedge over the loci it contains, following the hypergraph treatment of multi-way genome contacts of Dotson et al. [1]. Everything that follows is built on this object: two geometric scores, a partition of the reads into two classes, and a means of tuning the contact ratio by mixing those classes.

2. Data and Representation

The data is Pore-C from mouse embryonic stem cells, restricted to a core of 195 loci across 18 chromosomes and 29,588 reads. It is encoded as a binary incidence matrix $B \in \{0,1\}^{195\times 29{,}588}$, where $B_{ij} = 1$ when locus i appears in read j. Reads span between two and eight loci. A read confined to one chromosome is intra-chromosomal; a read touching two or more is inter-chromosomal. Population-wide, the two run about 5.5 : 1 in favour of intra, the skewed baseline that Section 4 sets out to shift.

Nothing else enters. No coordinates, sequence, or annotation informs the two scores; B carries all the structure, and every quantity below is derived from it.

3. Methods

3.1 Nonlinear eigenvector centrality

The first score rests on a deliberately circular definition: a locus is important if it belongs to important reads, and a read is important if it joins important loci. The circularity is resolved by iteration. Two vectors carry the scores, x over loci and y over reads. The scheme is the node–edge nonlinear centrality of Tudisco and Higham [2], in the form Lawson et al. [3] apply to hypernetworks. With W and N diagonal weight matrices on edges and nodes (here the identity), one step reads

$$\begin{aligned} u &= \sqrt{x \odot g\big(B\,W f(y)\big)}, & v &= \sqrt{y \odot \psi\big(B^{\top} N \varphi(x)\big)}, \\ x &\leftarrow u/\|u\|_1, & y &\leftarrow v/\|v\|_1. \end{aligned} \tag{1}$$

The two matrix products carry the propagation. $B\,(W f(y))$ gives each locus the summed, f-transformed score of the reads it belongs to; $B^{\top}(N\varphi(x))$ gives each read the summed, φ-transformed score of the loci it contains. The remaining operations stabilize the iteration: the square root blends the new estimate with the previous one, preventing oscillation, and the ℓ_1 normalization holds the two sides on a common scale. On convergence the pair (x^*, y^*) is stationary, and that fixed point is the centrality.

The four functions f, g, φ, ψ constitute the model, and they act in a single place: the manner in which a read aggregates the scores of its loci, $\psi\big(\sum_i \varphi(x_i)\big)$. This expression is a generalized mean, and three standard choices yield three behaviours:

- **Linear** (every function the identity): the read takes the **sum**, so a single dominant locus can determine it. This is ordinary linear eigenvector centrality.

- **LogExp** ($g = \sqrt{\cdot}$, $\varphi = \log$, $\psi = \exp$): the read takes the **product**, the geometric mean, so it is large only when *every* locus is. This is the model used here.
- **Max** ($\varphi = x^{\alpha}$, $\psi = x^{1/\alpha}$, large α): the read retains its **largest** locus and discards the rest.

The rationale for LogExp is that a product remains small unless all of a read's loci are large, which is precisely the property a multi-way contact should reward: the collective importance of the whole set, not of a single member. The sum is too permissive, since one large locus raises the read irrespective of the others. The maximum is too severe, retaining only one locus. LogExp is also well-behaved analytically: within the Tudisco–Higham framework the map is a contraction in the Hilbert metric with a unique positive fixed point, and the log–exp pair lies well inside the regime where this holds, whereas the Max model approaches its boundary as α grows. What is missing is the empirical test. The three models have not been run on this dataset and compared by the cores they produce, so the claim that LogExp separates this data best remains to be established rather than granted by the framework.



Figure 2: *The three centrality models as a spectrum of aggregations: arithmetic (Linear), geometric / soft-max (LogExp), and winner-take-all (Max).*

3.2 Forman–Ricci curvature on the poset

Curvature is the second score, and it carries the central argument of the paper. Geometrically, it measures how a connection is situated among its neighbours: one embedded in a dense region of overlapping structure is curved in one sense, one that spans a gap in the other. To make this precise on a hypergraph, we first recast the hypergraph as a poset. Augment the incidence matrix as $\tilde{H} = [\, I \mid H \,]$, so that loci and reads become elements of a single ordered set. Order those elements by size, and draw a covering relation wherever one is directly contained in another; this gives the Hasse diagram A. Finally, the 2-simplices (triangles) are obtained from $A \cdot A$, marking where elements share a higher-order face.

Forman–Ricci curvature [4] then scores each poset edge (i, j) as

$$R_F(i,j) \;=\; 4 \;-\; \deg(i) \;-\; \deg(j) \;+\; 3\,T_{ij}, \tag{2}$$

with $\deg(\cdot)$ the Hasse-diagram degrees and T_{ij} the number of triangles carried by the edge. The two terms oppose one another. Degree lowers the curvature, since an element with many parallel neighbours disperses structure in the manner of a bridge, while triangles raise it, since shared higher-order faces are the signature of a tight cluster. Aggregated to each read and taken in absolute, normalized form, this yields the per-edge score.

This score defines the two classes on which the remainder of the paper depends. Reads whose curvature is large in magnitude, those embedded most deeply in local structure, are termed **interior**;

reads whose curvature is small, the geometrically flat ones, are termed **bridge**. The terms reflect the intuition: an interior read lies within a structurally dense group of contacts, whereas a bridge read spans between such groups. One caution belongs here rather than in a later section. The split is taken on the *absolute* curvature, whereas the standard reading of a bridge, a connector between communities, corresponds specifically to *negative* curvature; the sign convention should therefore be verified before the signed interpretation is asserted. The magnitude split itself is not in doubt, nor is the empirical consequence established below: the interior reads are almost entirely inter-chromosomal.

3.3 A worked example

Both scores are clearest on a minimal example. Take three loci A, B, C and two reads, $e_1 = \{A, B, C\}$ and $e_2 = \{B, C\}$. Here e_1 is a three-way contact that no ordinary graph edge can represent, which is precisely the reason for hyperedges. The incidence matrix is

$$B = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{cc} e_1 & e_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \end{array}.$$

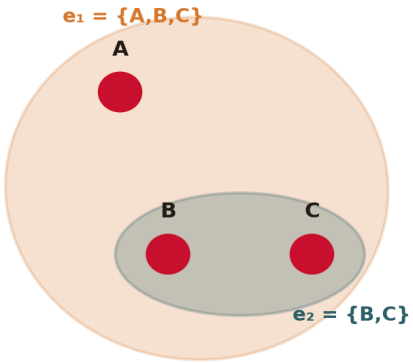


Figure 3: *A toy hypergraph: three loci and two hyperedges, $e_1 = \{A, B, C\}$ and $e_2 = \{B, C\}$.*

Centrality. Begin with all scores equal, $x = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $y = (\frac{1}{2}, \frac{1}{2})$, and perform one iteration of the Linear model. On the locus side $B\,y = (0.5, 1.0, 1.0)$, since A belongs to one read and B and C to two. Multiplying by the previous x, taking the square root, and normalizing gives $x \approx (0.26, 0.37, 0.37)$. On the read side $B^\top x = (1.0, 0.67)$, since e_1 holds three loci and e_2 two; the same operations give $y \approx (0.55, 0.45)$. The ordering is already correct after a single iteration: A ranks below B and C, and the three-way e_1 ranks above the two-way e_2; further iterations only sharpen it. Under LogExp the read update would multiply its loci's scores rather than sum them, so a read rises only when all of its loci do.

Curvature. Ordering the five elements by containment (Figure 4) places A below e_1, B and C below e_2, and e_2 below e_1. Since B and C reach e_1 only through e_2, those relations are not direct and do not appear. The degrees are $\deg(A) = \deg(B) = \deg(C) = 1$, $\deg(e_2) = 3$, and $\deg(e_1) = 2$, and this minimal example contains no triangles. Substituting into $R_F = 4 - \deg(i) - \deg(j) + 3T$ gives $+1$ on $A{-}e_1$, 0 on $B{-}e_2$ and $C{-}e_2$, and -1 on $e_2{-}e_1$. The high-degree element e_2 drives its

own links negative, illustrating the degree term; in real data, where reads overlap extensively, the triangle term becomes active and produces the strongly curved interior reads.

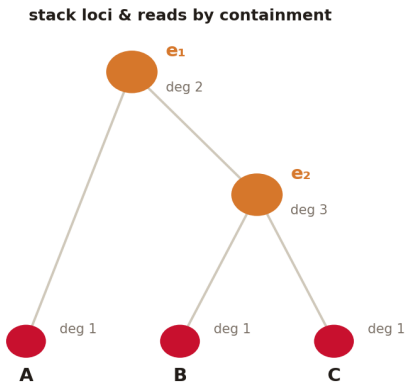


Figure 4: *The containment poset (Hasse diagram) of the toy hypergraph, with the degree of each element. Curvature is computed on this structure.*

3.4 Edge classes and the mixing scheme

Applying the curvature split to the data yields two classes of very unequal size. The interior class is small (214 reads) and almost purely inter-chromosomal, near 98%. The bridge class comprises the remainder, roughly 29,000 reads, and is about 85% intra-chromosomal. The inter-chromosomal signal is therefore not diffuse; it is concentrated in a small, sharply defined subset. Because the two pools have nearly opposite composition, any mixture attains an intra-to-inter ratio between theirs, and the mixing fraction moves that ratio smoothly across the range.

Two parameters set the mixture: ρ, the fraction of the interior pool retained, and κ, the fraction of the bridge pool retained. Drawing representative samples from two fixed pools, rather than thresholding a score, makes the resulting ratio a clean function of ρ and κ. This is the central step of the approach. The curvature has already separated the reads into an inter-rich pool and an intra-rich pool, so setting a contact ratio reduces to choosing how much of each to include.

4. Results

4.1 Two classes of opposite composition

Figure 5 shows the asymmetry on which the approach rests. The interior class is almost exclusively inter-chromosomal, while the far larger bridge class is predominantly intra-chromosomal. The contacts of interest are thus confined to a small, well-defined subset of the data.

4.2 Continuous tuning of the contact ratio

Retaining every interior read and adding an increasing fraction of the bridge pool raises the intra-to-inter ratio smoothly from inter-rich to intra-dominated, without discontinuity. A $3:1$ target is reached at roughly 5% of the bridge pool, about 1,700 reads in total. Any attainable ratio is set by a single bridge fraction.

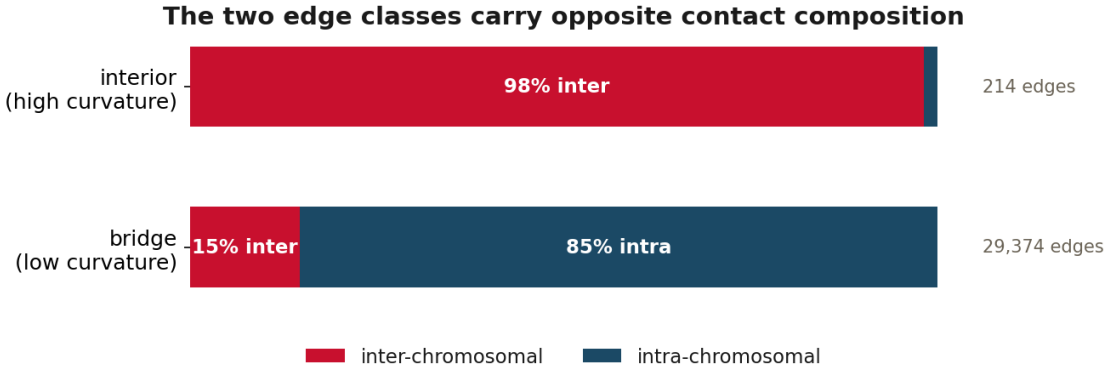


Figure 5: *Composition of the two edge classes. Interior edges are ∼98% inter-chromosomal; bridge edges are ∼85% intra-chromosomal.*

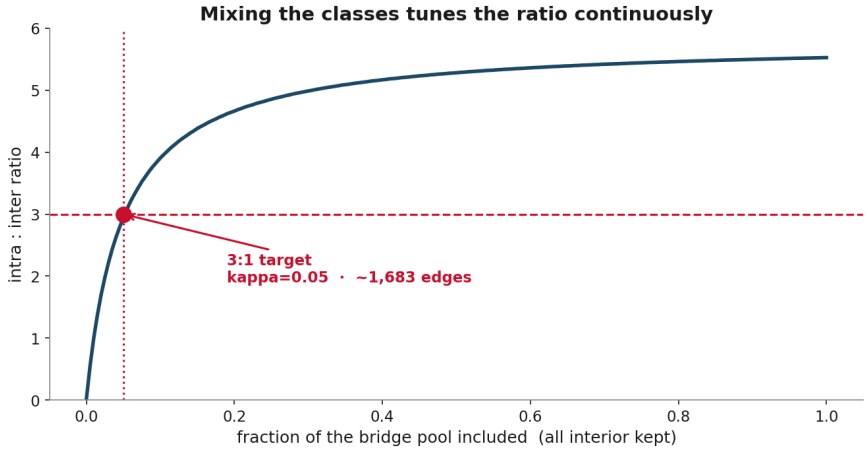


Figure 6: *The intra-to-inter ratio as a function of the bridge fraction added to the interior pool. The curve is smooth and monotone; a* $3{:}1$ *ratio is reached at about 5% of the bridge pool.*

4.3 The chromosome-pair graph and its stability

The inter-chromosomal reads of any mixture define a chromosome-by-chromosome graph whose off-diagonal cells count the reads linking two chromosomes (Figure 7). This graph is the readout for territory structure. Notably, it changes little: the dominant structure persists whether the mixture is inter-rich or intra-dominated, so the mixing parameter adjusts the intensity of the map rather than its form.

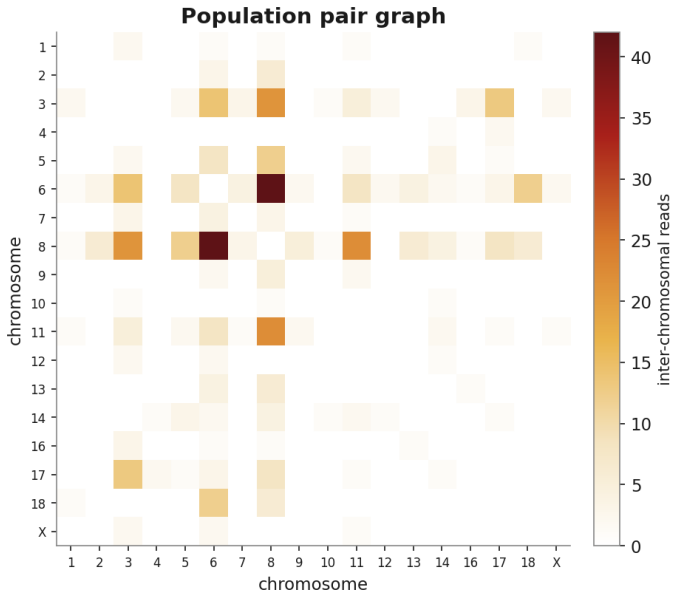


Figure 7: *The population chromosome-pair graph, aggregating inter-chromosomal co-occurrence across the dataset. Off-diagonal cells count reads linking the two chromosomes.*

4.4 A persistent chromosome-8 hub

One feature persists at every ratio: chromosome 8 lies at the centre of the inter-chromosomal map, with an especially strong contact to chromosome 6 (Figure 8). Because it holds across the entire sweep, it is not an artifact of a particular mixture. Whether it is genuine is a separate question, and it marks the boundary of what the result establishes. The hub may reflect genuine *trans* organization in these cells, or it may be an artifact of which loci entered the 195-locus core; only independent data can distinguish the two.

5. Discussion

The method's appeal is that ratio control is smooth and follows directly from quantities the hypergraph model already produces, with no thresholds to tune, and that the pair graph retains its form across the entire ratio range, which is the principal reason to regard the recovered structure as more than an artifact of a single mixture.

The limitations are worth stating directly. With only 195 loci over 18 chromosomes, the pair graph is sparse by construction, and many chromosome pairs cannot appear at all. The chromosome-8 hub is unverified. The analysis concerns a single population, and the mixing draws representative samples from two fixed pools rather than resting on a formalized sampling model. Two open points concern the method itself: the curvature split operates on absolute values, so the signed

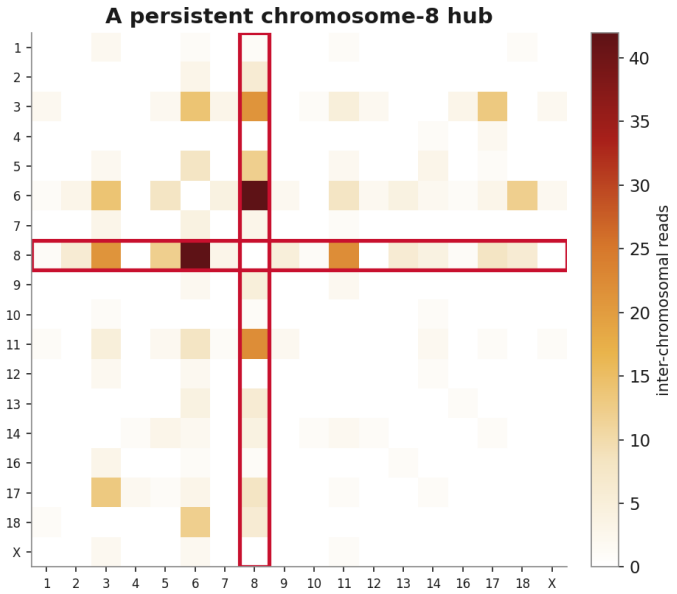


Figure 8: *The inter-chromosomal map with chromosome 8 highlighted. The chromosome-8 hub persists across all mixing ratios; whether it is biological or a sampling artifact is not yet established.*

interior/bridge interpretation requires confirmation, and the LogExp choice, well-motivated as it is, has not been compared against Linear and Max on this data.

The next steps follow from these gaps. The chromosome-8 hub should be validated against an independent measurement. The read scores should be related to the epigenetic tracks already present in the data object (accessibility, CTCF, H3K27ac, PolII), which is the most promising route to explaining why particular contacts are distinguished. Beyond that, the analysis should be extended past the 195-locus core, the mixing rule formalized, and the three-model comparison described in Section 3.1 carried out.

References

[1] Gabrielle A. Dotson, Can Chen, Stephen Lindsly, Anthony Cicalo, Sam Dilworth, Charles Ryan, Sivakumar Jeyarajan, Walter Meixner, Cooper Stansbury, Joshua Pickard, Nicholas Beckloff, Amit Surana, Max Wicha, Lindsey A. Muir, and Indika Rajapakse. Deciphering multi-way interactions in the human genome. *Nature Communications*, 13:5498, 2022.

[2] Francesco Tudisco and Desmond J. Higham. Node and edge nonlinear eigenvector centrality for hypergraphs. *Communications Physics*, 4:201, 2021.

[3] Sinan Lawson, Diane Donovan, and James Lefevre. An application of node and edge nonlinear hypergraph centrality to a protein complex hypernetwork. *PLOS ONE*, 19(10):e0311433, 2024.

[4] Robin Forman. Bochner's method for cell complexes and combinatorial ricci curvature. *Discrete & Computational Geometry*, 29(3):323–374, 2003.