Principal Curve approaches for inferring 3D Chromatin Architecture

Elena Tuzhilina

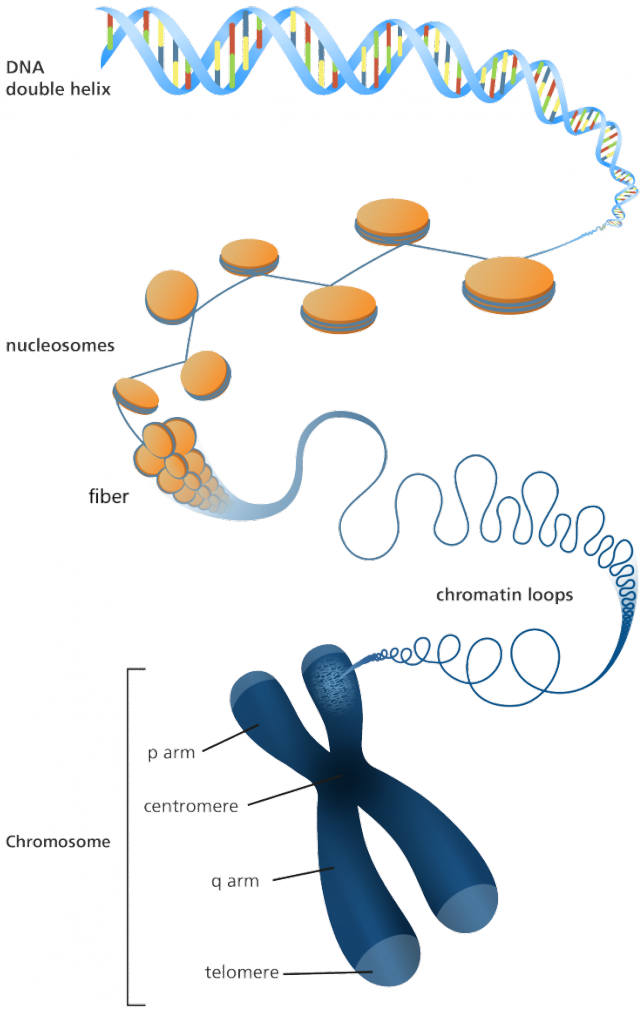
(joint work with Trevor Hastie and Mark Segal)

Stanford University, Department of Statistics

elenatuz@stanford.edu

August 6, 2020

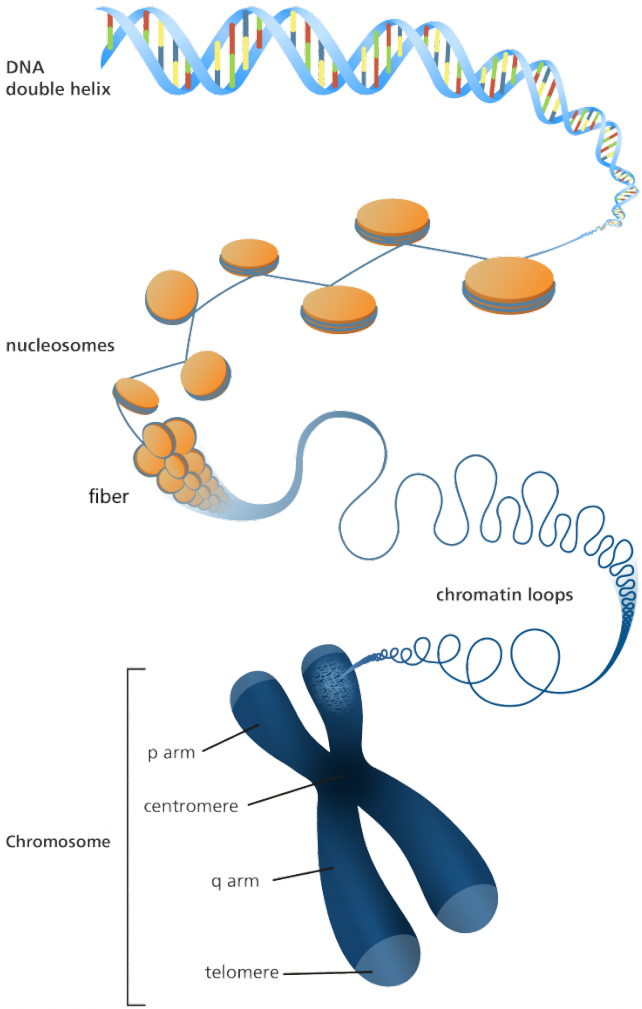
Chromatin



Chromatin = DNA + nucleosomes

1. DNA

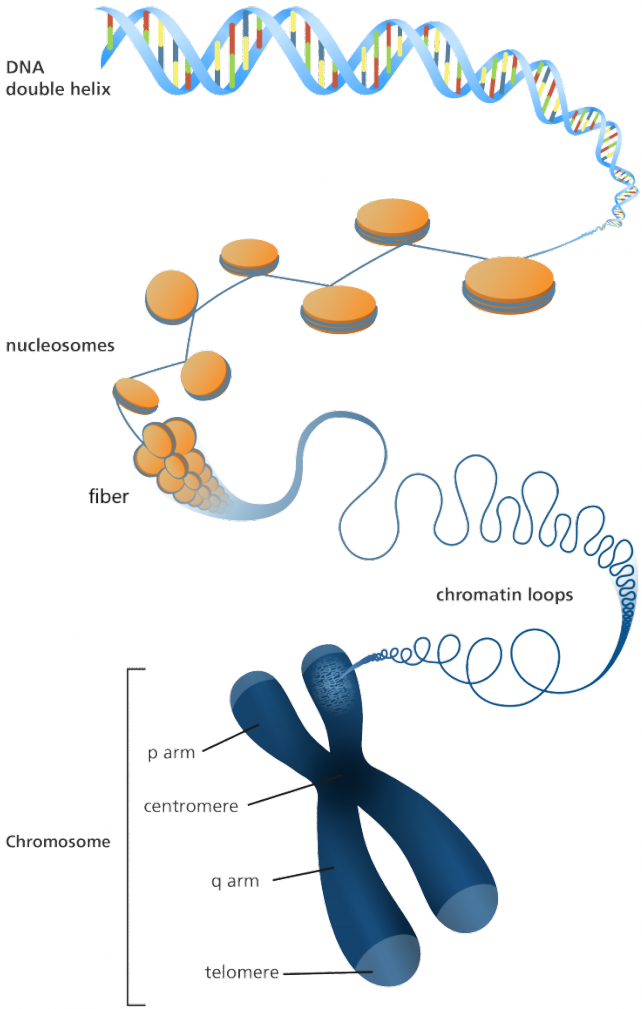
Chromatin



Chromatin = DNA + nucleosomes

1. DNA
2. 'Beads-on-a-string'

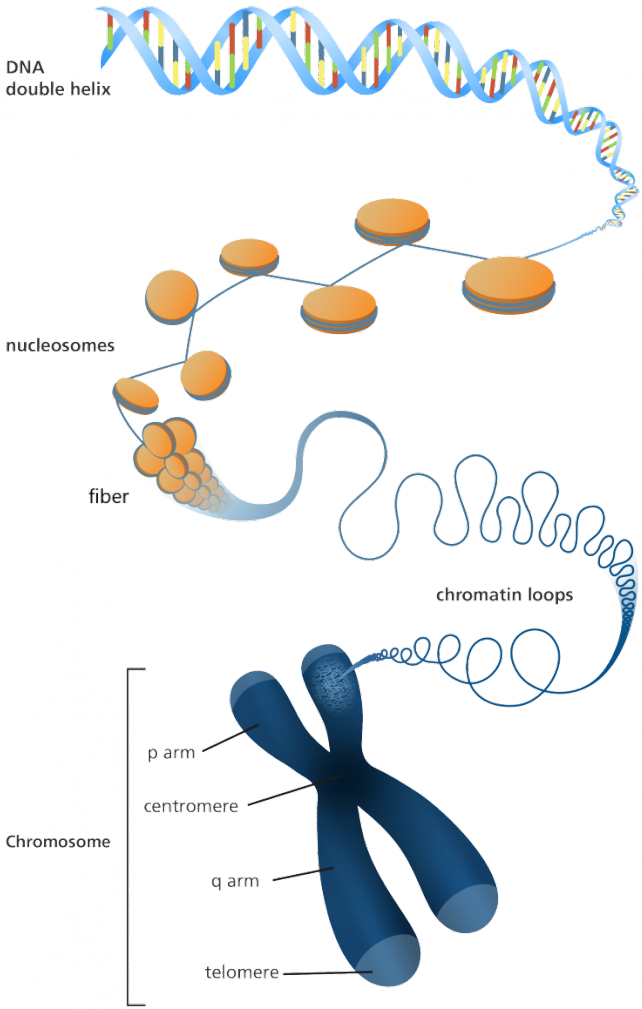
Chromatin



Chromatin = DNA + nucleosomes

1. DNA
2. 'Beads-on-a-string'
3. Chromatin fiber

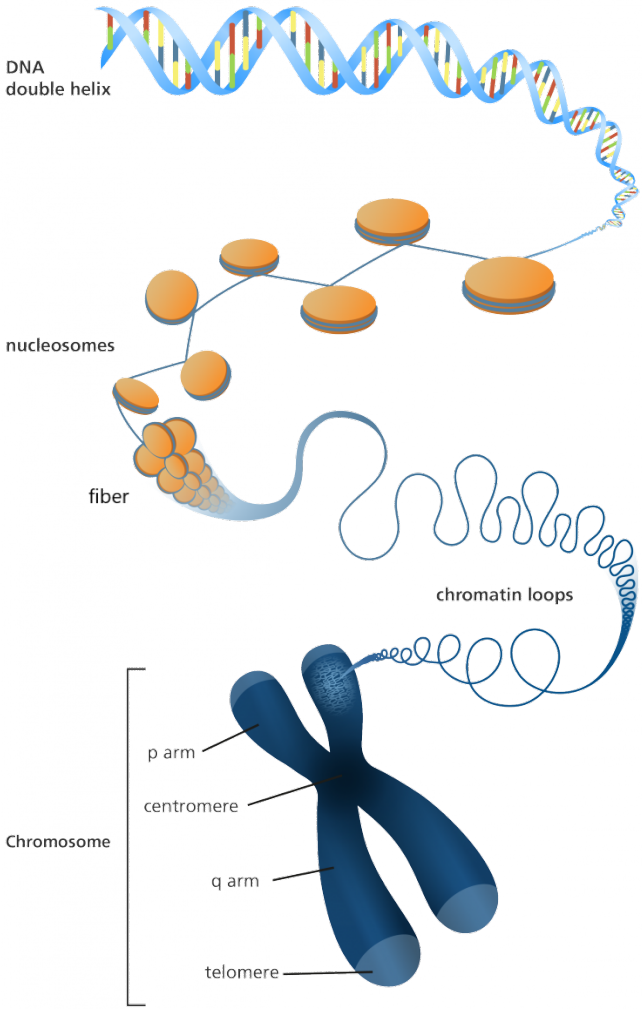
Chromatin



Chromatin = DNA + nucleosomes

1. DNA
2. 'Beads-on-a-string'
3. Chromatin fiber
4. Chromatin loop

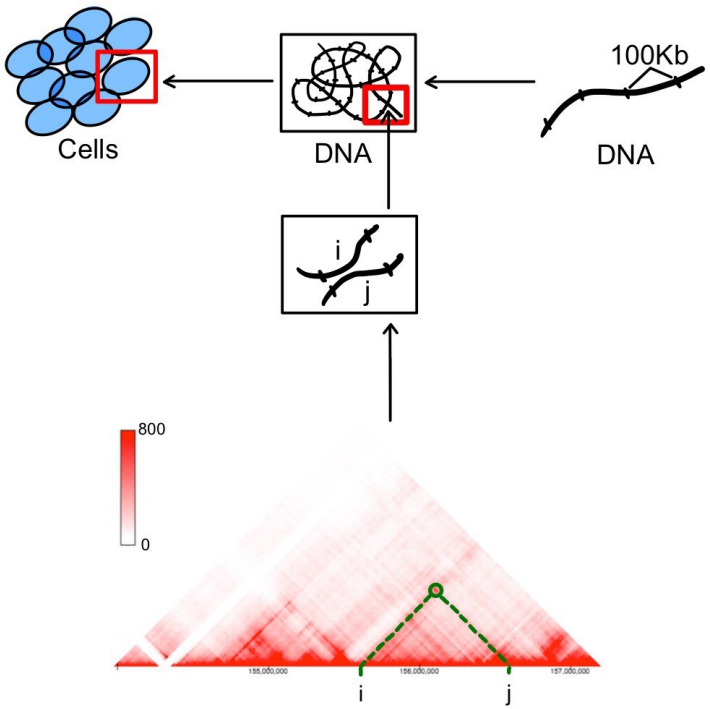
Chromatin



Chromatin = DNA + nucleosomes

1. DNA
2. 'Beads-on-a-string'
3. Chromatin fiber
4. Chromatin loop
5. Chromosome

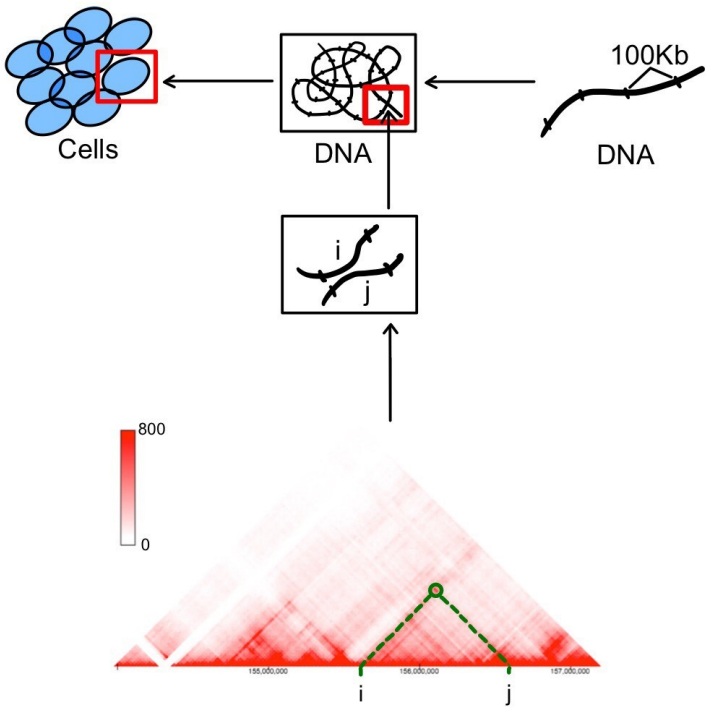
Contact matrix



Terminology

- genomic locus = 'piece'
- resolution = 'size of a piece'
- contact (formaldehyde + cross-linking + sequencing)

Contact matrix



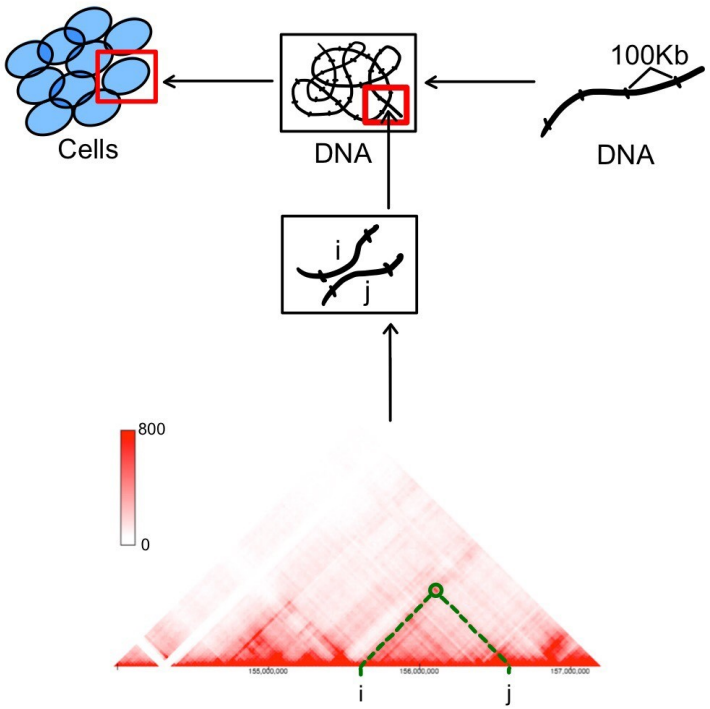
Terminology

- genomic locus = 'piece'
- resolution = 'size of a piece'
- contact (formaldehyde + cross-linking + sequencing)

Notations

- $n = \#$ genomic loci
- $C_{ij} = \#$ contacts between loci i and j

Contact matrix



Terminology

- genomic locus = 'piece'
- resolution = 'size of a piece'
- contact (formaldehyde + cross-linking + sequencing)

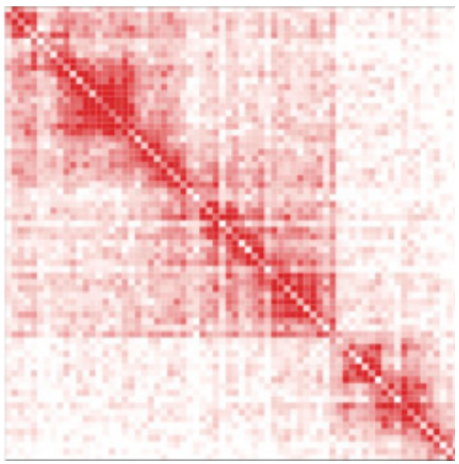
Notations

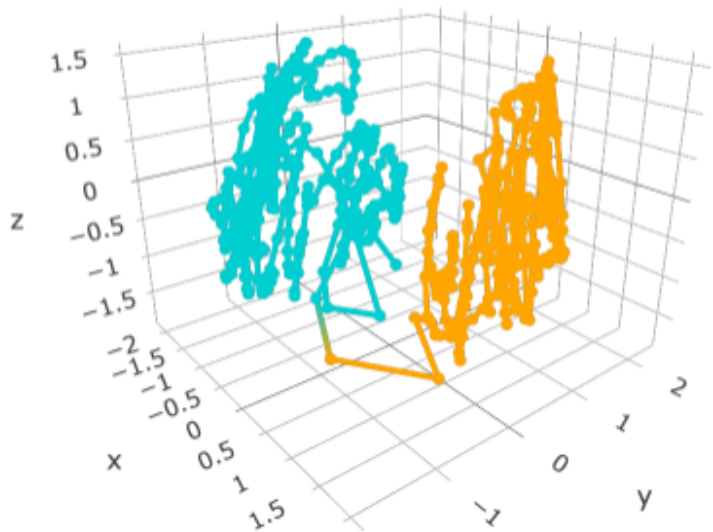
- $n = \#$ genomic loci
- $C_{ij} = \#$ contacts between loci i and j

Contact matrix $C = [C_{ij}] \in \mathbb{Z}_+^{n \times n}$

Chromatin reconstruction problem

Goal: Use the information contained in C to reconstruct the locus spatial coordinates $x_1, \ldots, x_n \in \mathbb{R}^3$.





Examples

Main ingredients

- loss function $\ell(x_1, \ldots, x_n)$
- optimization problem minimizing/maximizing $\ell(x_1, \ldots, x_n)$ w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Examples

Main ingredients

- loss function $\ell(x_1, \ldots, x_n)$
- optimization problem minimizing/maximizing $\ell(x_1, \ldots, x_n)$ w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Example (Deterministic model, Metric MDS)

1. Convert C to a distance matrix D, e.g. $D_{ij} = \begin{cases} (C_{ij})^{-\alpha} \text{ if } C_{ij} > 0 \\ \infty \text{ if } C_{ij} = 0 \end{cases}$
2. Minimize Stress objective

$$\ell(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} W_{ij} \left(D_{ij} - \|x_i - x_j\|\right)^2$$

w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Examples

Main ingredients

- loss function $\ell(x_1, \ldots, x_n)$
- optimization problem minimizing/maximizing $\ell(x_1, \ldots, x_n)$ w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Example (Probability model, Poisson)

1. $C_{ij} \sim Pois(\lambda_{ij})$, where $\lambda_{ij} = \lambda_{ij}(x_1, \ldots, x_n) = \beta \|x_i - x_j\|^\alpha$
2. Minimize negative log-likelihood

$$\ell(x_1, \ldots, x_n) = \sum_{1 \le i,j \le n} \beta \|x_i - x_j\|^\alpha - C_{ij} \log\left(\beta \|x_i - x_j\|^\alpha\right)$$

w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Controlling reconstruction smoothness

(Previous approaches) Add a smoothness penalty!

$$\text{minimize } \ell(x_1, \ldots, x_n) + \lambda h(x_1, \ldots, x_n)$$
$$\text{w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3$$

Controlling reconstruction smoothness

(Previous approaches) Add a smoothness penalty!

$$\text{minimize } \ell(x_1, \ldots, x_n) + \lambda h(x_1, \ldots, x_n)$$
$$\text{w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3$$

Problems

- non convexity
- complexity of the objective, gradient and hessian

Controlling reconstruction smoothness

(Previous approaches) Add a smoothness penalty!

$$\text{minimize } \ell(x_1, \ldots, x_n) + \lambda h(x_1, \ldots, x_n)$$
$$\text{w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3$$

Problems

- non convexity
- complexity of the objective, gradient and hessian

(Our approach) Add a constraint!

$$\text{minimize } \ell(x_1, \ldots, x_n) \text{ w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3$$
$$x_1, \ldots, x_n \in \text{ smooth one-dimensional curve}$$

Smooth curve constraint

1. $x_1, \ldots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j} \; h_\ell(t)$

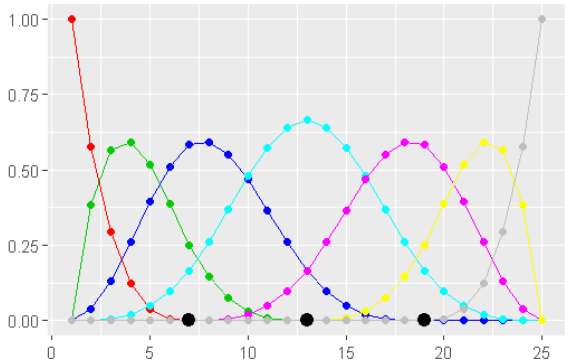
k = spline degrees of freedom (df), hyperparameter

Smooth curve constraint

1. $x_1, \ldots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \ldots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$



$$\longrightarrow \quad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$

$$X = \begin{pmatrix} -x_1^T - \\ \dots \\ -x_n^T - \end{pmatrix} = \begin{pmatrix} | & | & | \\ \gamma_1 & \gamma_2 & \gamma_3 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{n \times 3} \qquad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$

$$X = \begin{pmatrix} -x_1^T - \\ \dots \\ -x_n^T - \end{pmatrix} = \begin{pmatrix} | & | & | \\ \gamma_1 & \gamma_2 & \gamma_3 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{n \times 3} \qquad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

$x_1, \dots, x_n \in$ smooth one-dimensional curve $\iff$

$\exists\ \Theta \in \mathbb{R}^{k \times 3}$ such that $X = H\Theta$

PCMS = Classical MDS + Smooth curve constraint

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} (C_{ij} - \langle x_i, \ x_j \rangle)^2 \Longleftrightarrow \ell(X) = \| C - XX^T \|_F^2$$

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} (C_{ij} - \langle x_i, \ x_j \rangle)^2 \Longleftrightarrow \ell(X) = \| C - XX^T \|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - \langle x_i,\ x_j\rangle)^2 \Longleftrightarrow \ell(X) = \|C - XX^T\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|C - H\Theta\Theta^T H^T\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k\times 3}$$

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - \langle x_i,\ x_j\rangle)^2 \Longleftrightarrow \ell(X) = \|C - XX^T\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|C - H\Theta\Theta^T H^T\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k\times 3}$$

Solution via Eigen Decomposition of $H^T CH$

PCMS + weights

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(C_{ij} - \langle x_i,\, x_j\rangle)^2 \Longleftrightarrow \ell(X) = \|\sqrt{W} * (C - XX^T)\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|\sqrt{W} * (C - H\Theta\Theta^T H^T)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k\times 3}$$

Solution iterative algorithm using PCMS as a building block

PCMS examples

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

PCMS examples

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

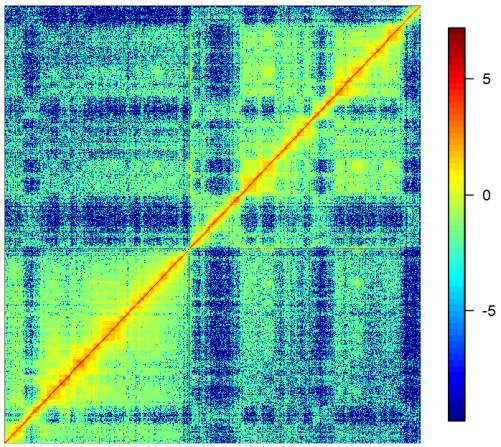
$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha}$ for $\epsilon = 0.001$, $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$

PCMS examples ($df = 10$)

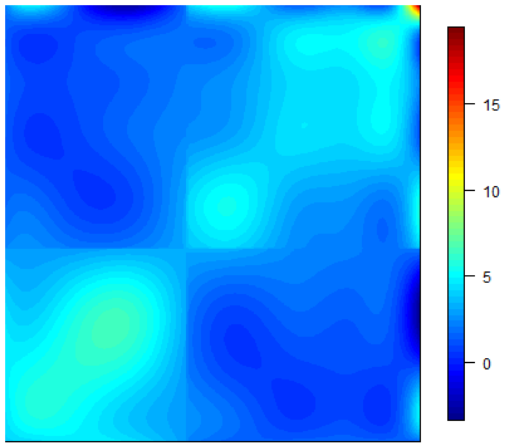
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

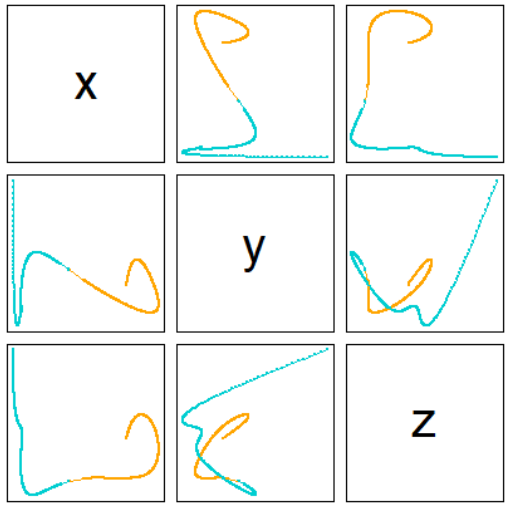
$$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha} \text{ for } \epsilon = 0.001,\ \alpha = 1 \text{ and } \beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$$



Original data



Reconstruction



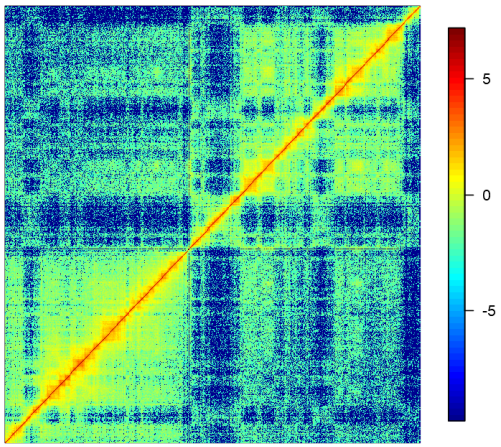
3D conformation

PCMS examples ($df = 25$)

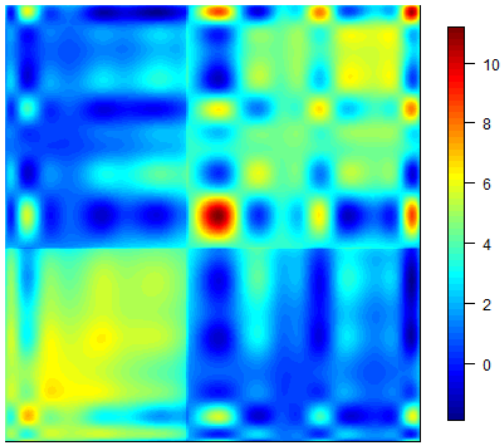
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

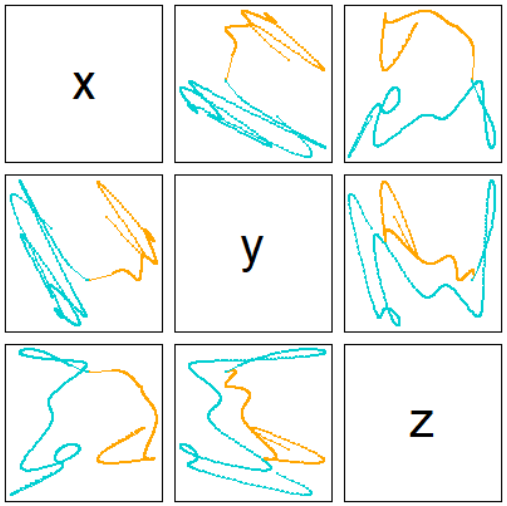
$$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha} \text{ for } \epsilon = 0.001,\ \alpha = 1 \text{ and } \beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$$



Original data



Reconstruction



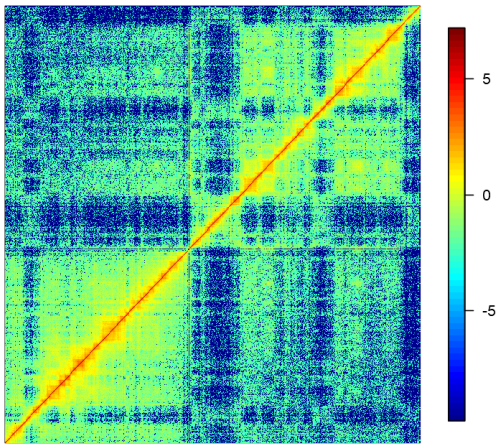
3D conformation

PCMS examples ($df = 50$)

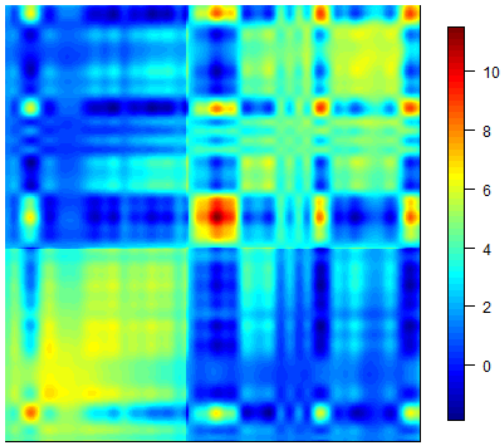
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

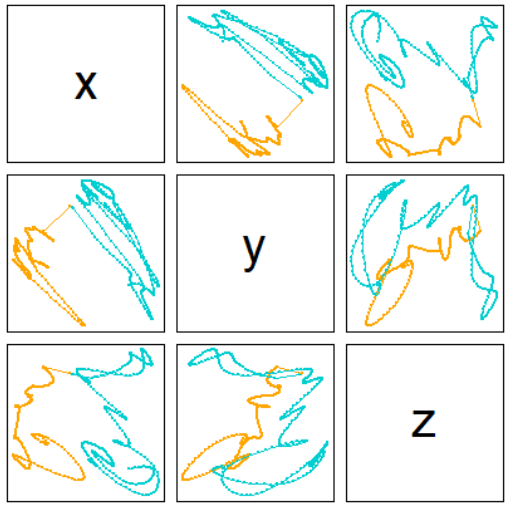
$$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha} \text{ for } \epsilon = 0.001,\ \alpha = 1 \text{ and } \beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$$



Original data



Reconstruction



3D conformation

PoisMS

PoisMS = Weighted PCMS + Poisson Model

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X) = \sum_{1\leq i,j\leq n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right]$$

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X) = \sum_{1\leq i,j\leq n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right]$$

Smooth curve constraint $X = H\Theta$

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X) = \sum_{1\leq i,j\leq n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right]$$

Smooth curve constraint $X = H\Theta$

PoisMS optimization problem

$$\text{minimize } \ell_{PoisMS}(X) \text{ w.r.t. } X \text{ subject to } X = H\Theta$$

PoisMS

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X) = \sum_{1\leq i,j\leq n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right]$$

Smooth curve constraint $X = H\Theta$

PoisMS optimization problem

$$\text{minimize } \ell_{PoisMS}(X) \text{ w.r.t. } X \text{ subject to } X = H\Theta$$

Solution use PCMS as a building block of the iterative algorithm!

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha} \frac{C - W}{W}$

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha} \frac{C - W}{W}$

$\Longrightarrow$ **can use Weighted PCMS to optimize SOA loss**

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha}\frac{C - W}{W}$

$\Longrightarrow$ **can use Weighted PCMS to optimize SOA loss**

1. **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k \times 3}$

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha} \frac{C - W}{W}$

$\Longrightarrow$ **can use Weighted PCMS to optimize SOA loss**

1. **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k \times 3}$
2. *Repeat until convergence*:

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha}\frac{C - W}{W}$

$\Longrightarrow$ **can use Weighted PCMS to optimize SOA loss**

1. **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k \times 3}$
2. *Repeat until convergence*:
   - **[SOA]** $\hat{C} = H\Theta\Theta^T H^T, \quad W = e^{\alpha \cdot \hat{C} + \beta}, \quad Z = \hat{C} + \frac{1}{\alpha}(\frac{C-W}{W})$

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha}\frac{C - W}{W}$

$\Longrightarrow$ **can use Weighted PCMS to optimize SOA loss**

1. **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k \times 3}$
2. *Repeat until convergence*:
   - **[SOA]** $\hat{C} = H\Theta\Theta^T H^T, \quad W = e^{\alpha \cdot \hat{C} + \beta}, \quad Z = \hat{C} + \frac{1}{\alpha}(\frac{C-W}{W})$
   - **[WPCMS]** $\Theta := \text{PCMS}_W(Z, H)$

PoisMS examples

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

PoisMS examples

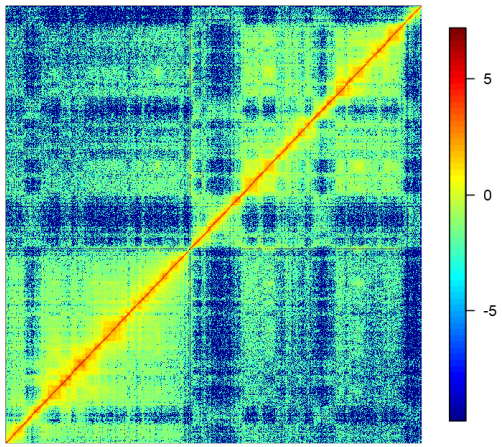
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$

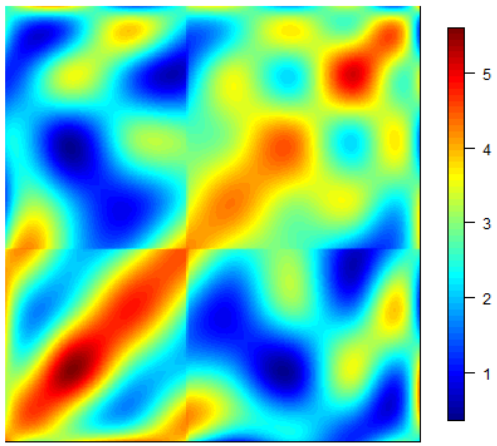
PoisMS examples ($df = 10$)

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

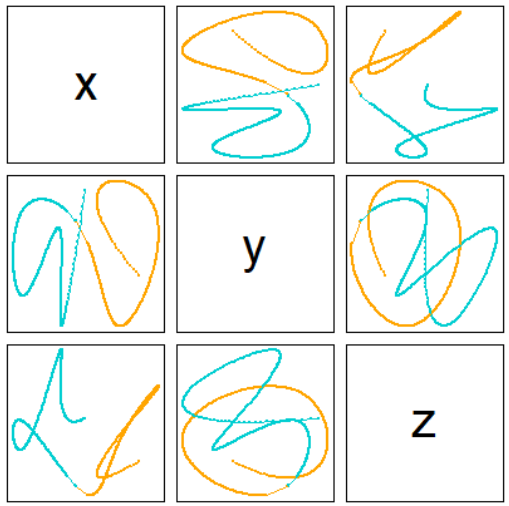
PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$



Original data



Reconstruction

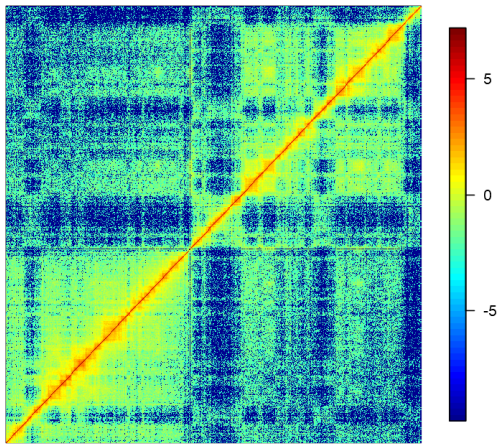


3D conformation

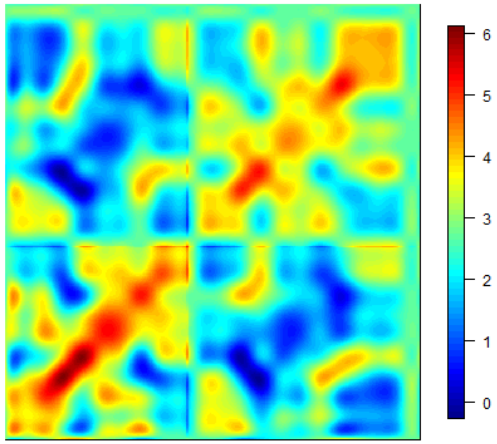
PoisMS examples ($df = 25$)

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

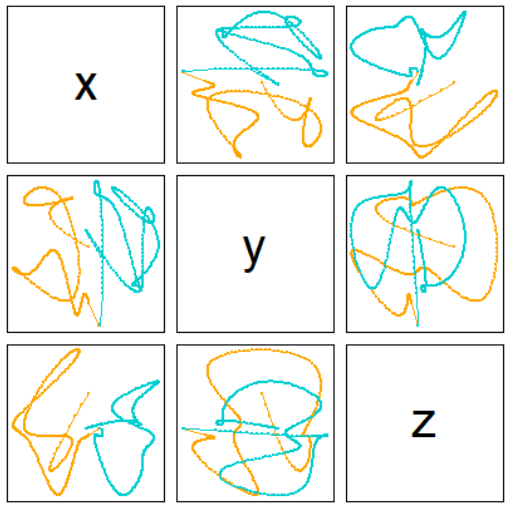
PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$



Original data



Reconstruction

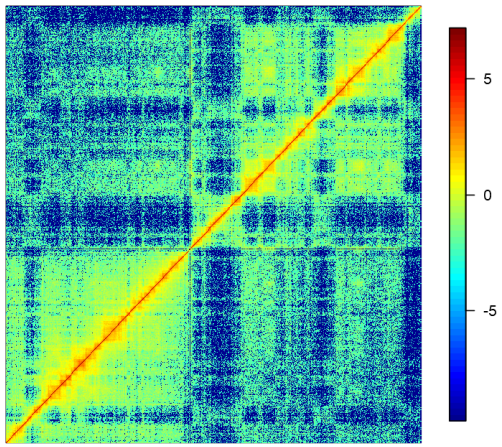


3D conformation

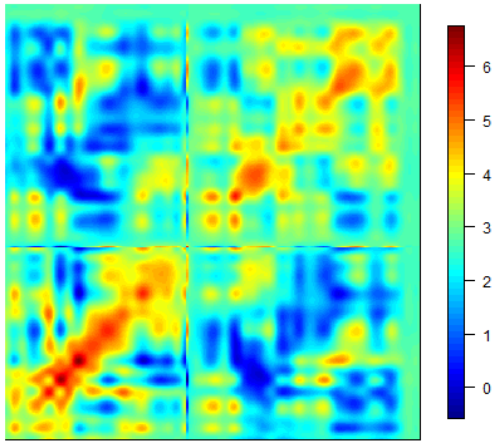
PoisMS examples ($df = 50$)

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

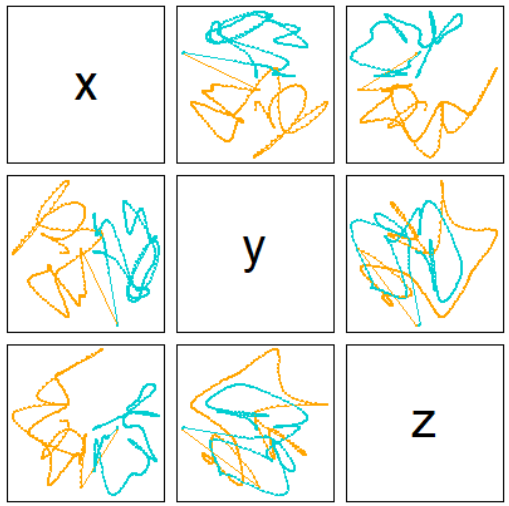
PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$



Original data



Reconstruction



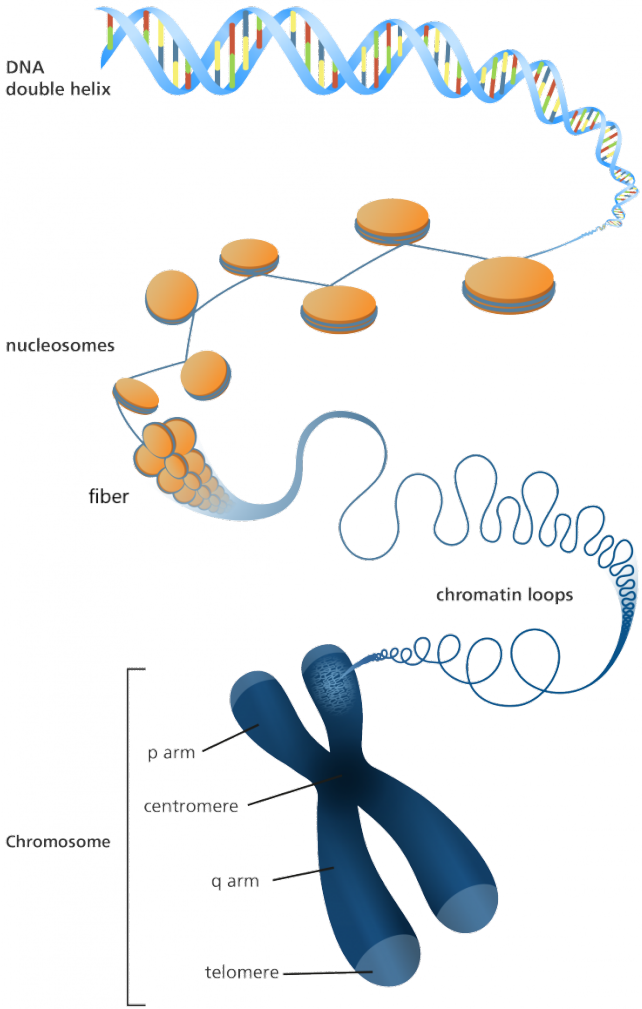
3D conformation

References

- T. Stevens et al. *3D structures of individual mammalian genomes studied by single-cell Hi-C*. Nature, 544:59–64, 2017.
- N. Varoquaux et al. *A statistical approach for inferring the three-dimensional structure of the genome*. Bioinformaics, 30(12):26-33, 2014.
- Z. Zhang et al. *3D Chromosome Modeling with Semi-Definite Programming and Hi-C Data*. Journal of computational biology, 20(11):831–846, 2013.
- C. Zou et al. *HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure*. Genome Biology, 17(40), 2016.
- M. Rosenthal et al. *Bayesian Estimation of 3D Chromosomal Structure from Single Cell Hi-C Data*. Journal of Computational Biology, published online, 2019.
- A. Buja et al. *Data Visualization With Multidimensional Scaling*. Journal of Computational and Graphical Statistics, 17(2):444-472, 2008.
- R. Mazumder et al. *Spectral Regularization Algorithms for Learning Large Incomplete Matrices*. Journal of Machine Learning Research, 11: 2287-2322, 2010.
- N. Srebro et al. *Weighted Low-Rank Approximations*. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

Thank you for your attention!

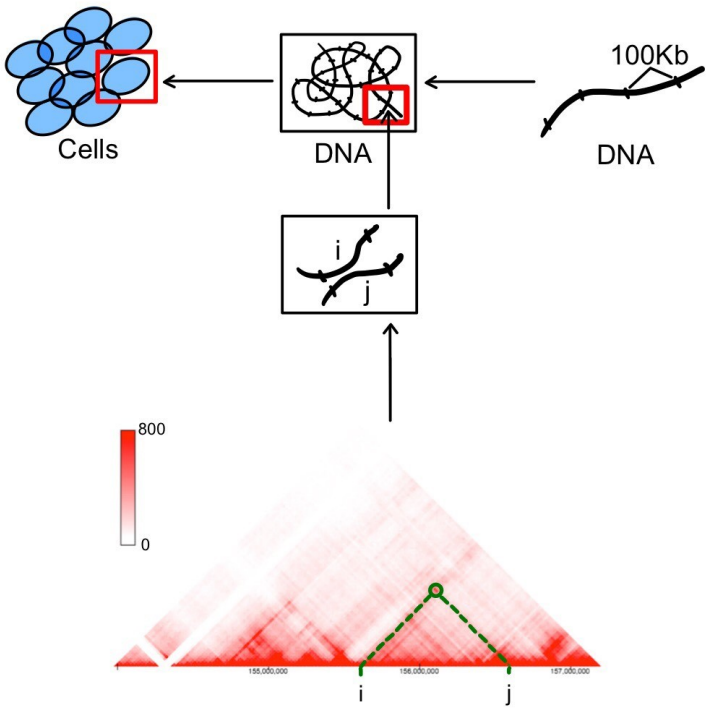
Chromatin



Chromatin = DNA + nucleosomes

1. DNA
2. 'Beads-on-a-string'
3. Chromatin fiber
4. Chromatin loop
5. Chromosome

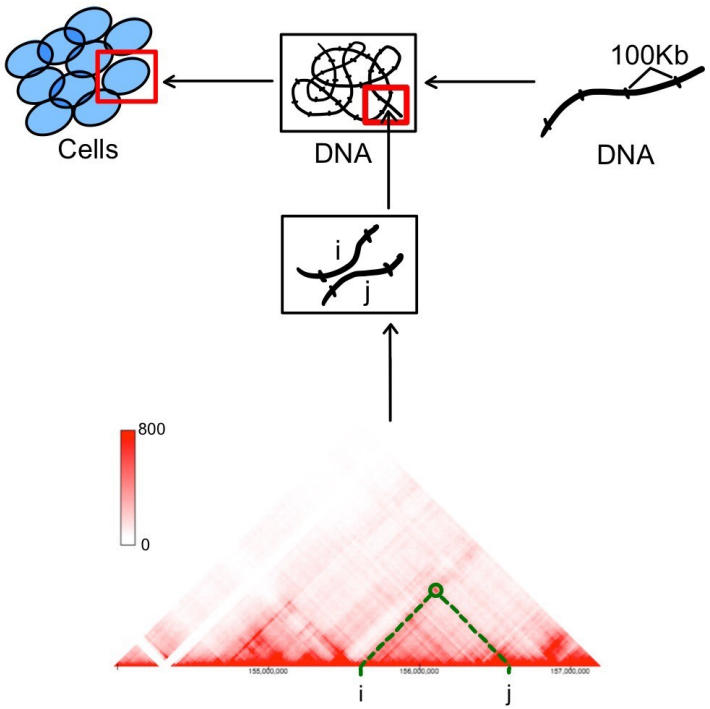
Contact matrix



Terminology

- genomic locus = 'piece'
- resolution = 'size of a piece'
- contact (formaldehyde + cross-linking + sequencing)

Contact matrix



Terminology

- genomic locus = 'piece'
- resolution = 'size of a piece'
- contact (formaldehyde + cross-linking + sequencing)

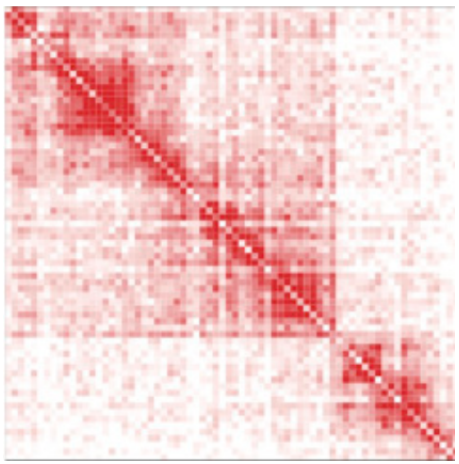
Notations

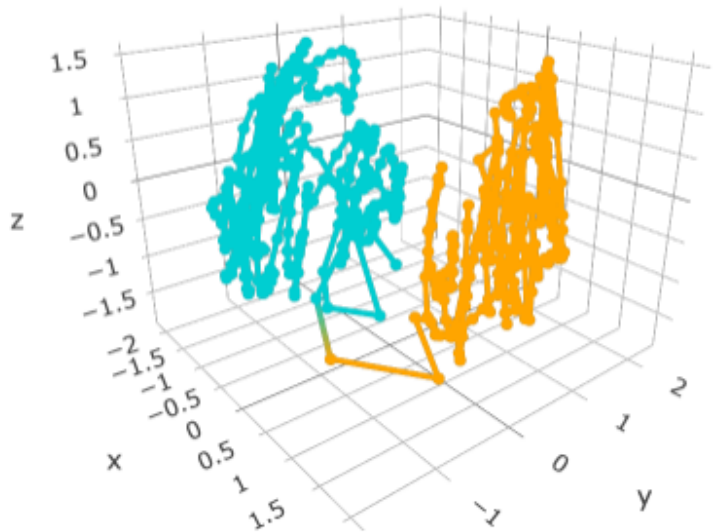
- $n = \#$ genomic loci
- $C_{ij} = \#$ contacts between loci i and j

Contact matrix $C = [C_{ij}] \in \mathbb{Z}_+^{n \times n}$

Chromatin reconstruction problem

Goal: Use the information contained in C to reconstruct the locus spatial coordinates $x_1, \ldots, x_n \in \mathbb{R}^3$.





Examples

Main ingredients

- loss function $\ell(x_1, \ldots, x_n)$
- optimization problem minimizing/maximizing $\ell(x_1, \ldots, x_n)$ w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Examples

Main ingredients

- loss function $\ell(x_1, \ldots, x_n)$
- optimization problem minimizing/maximizing $\ell(x_1, \ldots, x_n)$ w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Example (Deterministic model, Metric MDS)

1. Convert C to a distance matrix D, e.g. $D_{ij} = \begin{cases} (C_{ij})^{-\alpha} \text{ if } C_{ij} > 0 \\ \infty \text{ if } C_{ij} = 0 \end{cases}$
2. Minimize Stress objective

$$\ell(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} W_{ij} \left(D_{ij} - \|x_i - x_j\|\right)^2$$

w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Examples

Main ingredients

- loss function $\ell(x_1, \ldots, x_n)$
- optimization problem minimizing/maximizing $\ell(x_1, \ldots, x_n)$ w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Example (Probability model, Poisson)

1. $C_{ij} \sim Pois(\lambda_{ij})$, where $\lambda_{ij} = \lambda_{ij}(x_1, \ldots, x_n) = \beta \|x_i - x_j\|^\alpha$
2. Minimize negative log-likelihood

$$\ell(x_1, \ldots, x_n) = \sum_{1 \le i,j \le n} \beta \|x_i - x_j\|^\alpha - C_{ij} \log\left(\beta \|x_i - x_j\|^\alpha\right)$$

w.r.t. $x_1, \ldots, x_n \in \mathbb{R}^3$

Controlling reconstruction smoothness

(Previous approaches) Add a smoothness penalty!

$$\text{minimize } \ell(x_1, \ldots, x_n) + \lambda h(x_1, \ldots, x_n)$$
$$\text{w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3$$

Problems

- non convexity
- complexity of the objective, gradient and hessian

Controlling reconstruction smoothness

(Previous approaches) Add a smoothness penalty!

$$\text{minimize } \ell(x_1, \ldots, x_n) + \lambda h(x_1, \ldots, x_n)$$
$$\text{w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3$$

Problems

- non convexity
- complexity of the objective, gradient and hessian

(Our approach) Add a constraint!

$$\text{minimize } \ell(x_1, \ldots, x_n) \text{ w.r.t. } x_1, \ldots, x_n \in \mathbb{R}^3$$
$$x_1, \ldots, x_n \in \text{ smooth one-dimensional curve}$$

Smooth curve constraint

1. $x_1, \ldots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$

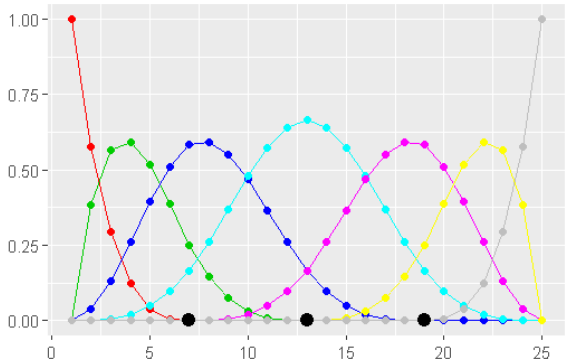
k = spline degrees of freedom (df), hyperparameter

Smooth curve constraint

1. $x_1, \ldots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \ldots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$



$$\longrightarrow \quad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$

$$X = \begin{pmatrix} -x_1^T - \\ \dots \\ -x_n^T - \end{pmatrix} = \begin{pmatrix} | & | & | \\ \gamma_1 & \gamma_2 & \gamma_3 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{n \times 3} \qquad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

Smooth curve constraint

1. $x_1, \dots x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
2. $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^{k} \Theta_{\ell j}\ h_\ell(t)$
3. $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^{k} \Theta_{\ell 1}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 2}\ h_\ell(i) \\ \sum_{\ell=1}^{k} \Theta_{\ell 3}\ h_\ell(i) \end{pmatrix}$

$$X = \begin{pmatrix} -x_1^T - \\ \dots \\ -x_n^T - \end{pmatrix} = \begin{pmatrix} | & | & | \\ \gamma_1 & \gamma_2 & \gamma_3 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{n \times 3} \qquad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

$x_1, \dots, x_n \in$ smooth one-dimensional curve $\iff$

$\exists\ \Theta \in \mathbb{R}^{k \times 3}$ such that $X = H\Theta$

PCMS

PCMS = Classical MDS + Smooth curve constraint

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - \langle x_i,\ x_j\rangle)^2 \Longleftrightarrow \ell(X) = \|C - XX^T\|_F^2$$

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - \langle x_i,\ x_j\rangle)^2 \Longleftrightarrow \ell(X) = \|C - XX^T\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|C - H\Theta\Theta^T H^T\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k\times 3}$$

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - \langle x_i,\ x_j\rangle)^2 \Longleftrightarrow \ell(X) = \|C - XX^T\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|C - H\Theta\Theta^T H^T\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k\times 3}$$

Solution via Eigen Decomposition of $H^T CH$

PCMS + weights

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(C_{ij} - \langle x_i,\, x_j\rangle)^2 \Longleftrightarrow \ell(X) = \|\sqrt{W} * (C - XX^T)\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|\sqrt{W} * (C - H\Theta\Theta^T H^T)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k\times 3}$$

Solution iterative algorithm using PCMS as a building block

PCMS examples

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

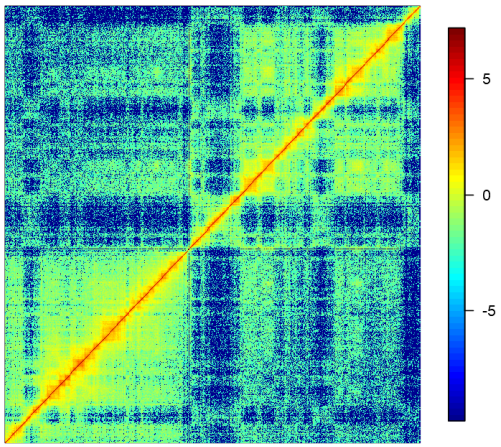
$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha}$ for $\epsilon = 0.001$, $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$

PCMS examples ($df = 10$)

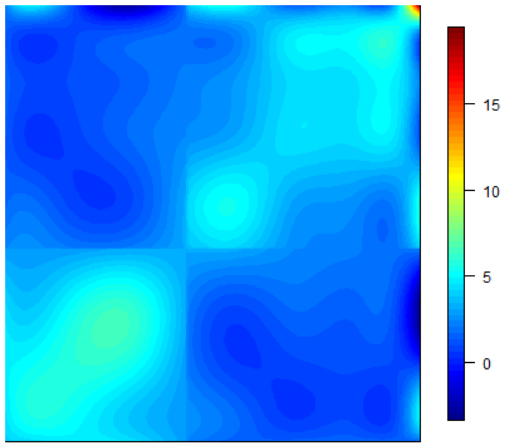
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

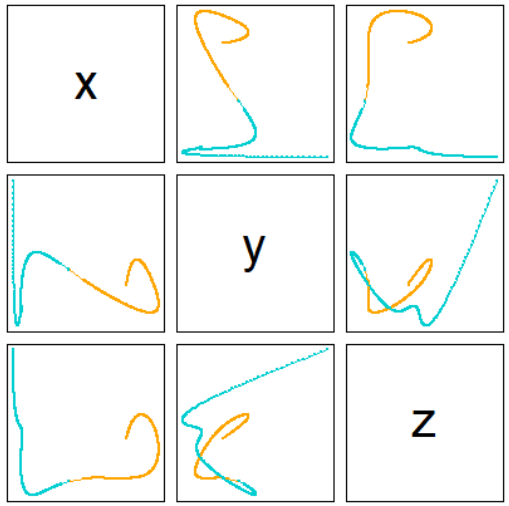
$$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha} \text{ for } \epsilon = 0.001,\ \alpha = 1 \text{ and } \beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$$



Original data



Reconstruction



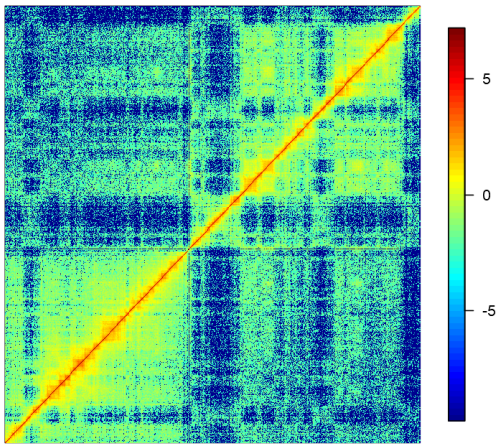
3D conformation

PCMS examples ($df = 25$)

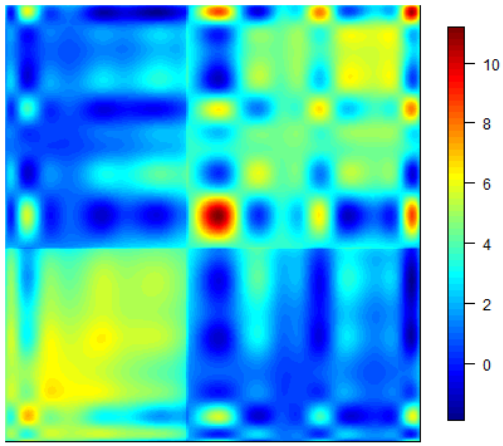
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

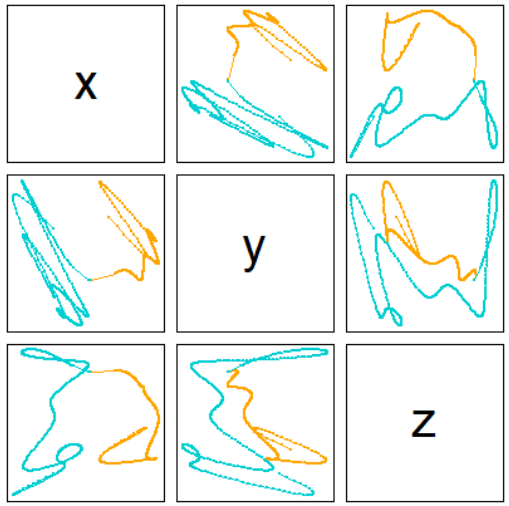
$$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha} \text{ for } \epsilon = 0.001,\ \alpha = 1 \text{ and } \beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$$



Original data



Reconstruction



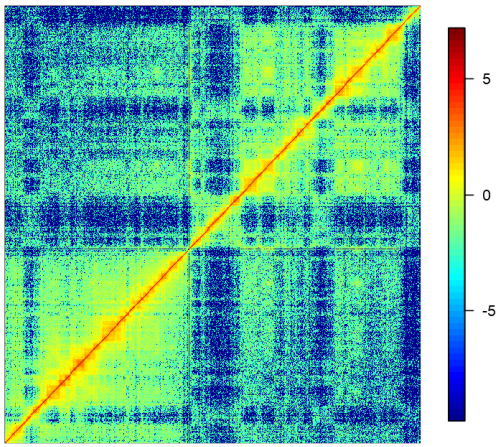
3D conformation

PCMS examples ($df = 50$)

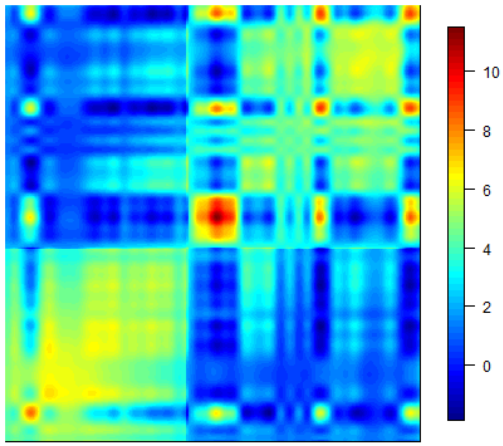
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

Transformation

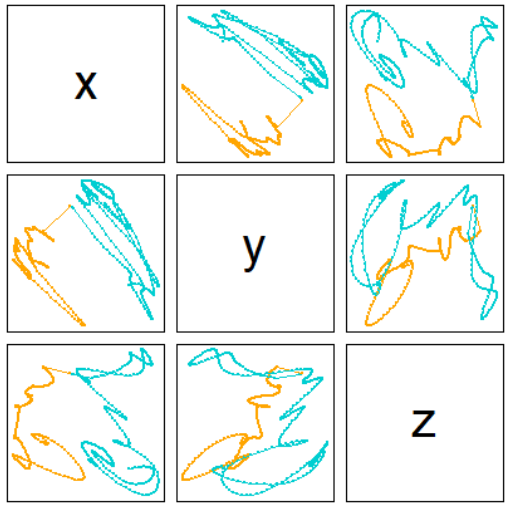
$$C^{\log} = \frac{\log(C+\epsilon)-\beta}{\alpha} \text{ for } \epsilon = 0.001,\ \alpha = 1 \text{ and } \beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$$



Original data



Reconstruction



3D conformation

PoisMS

PoisMS = Weighted PCMS + Poisson Model

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X) = \sum_{1\leq i,j\leq n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right]$$

PoisMS

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X) = \sum_{1\leq i,j\leq n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right]$$

Smooth curve constraint $X = H\Theta$

PoisMS optimization problem

$$\text{minimize } \ell_{PoisMS}(X) \text{ w.r.t. } X \text{ subject to } X = H\Theta$$

PoisMS

PoisMS = Weighted PCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X) = \sum_{1\leq i,j\leq n} \left[e^{\alpha\langle x_i, x_j\rangle + \beta} - C_{ij}\left(\alpha\langle x_i, x_j\rangle + \beta\right)\right]$$

Smooth curve constraint $X = H\Theta$

PoisMS optimization problem

$$\text{minimize } \ell_{PoisMS}(X) \text{ w.r.t. } X \text{ subject to } X = H\Theta$$

Solution use PCMS as a building block of the iterative algorithm!

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha} \frac{C - W}{W}$

$\Longrightarrow$ **can use Weighted PCMS to optimize SOA loss**

PoisMS iterative algorithm

$$\ell_{PoisMS}(X) \approx \ell_{SOA}(X) = \|\sqrt{W} * (Z - X^T X)\|_F^2$$

where $W = e^{\alpha X_0 X_0^T + \beta}$ and $Z = X_0 X_0^T + \frac{1}{\alpha}\frac{C - W}{W}$

$\Longrightarrow$ **can use Weighted PCMS to optimize SOA loss**

1. **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k \times 3}$
2. *Repeat until convergence*:
   - **[SOA]** $\hat{C} = H\Theta\Theta^T H^T, \quad W = e^{\alpha \cdot \hat{C} + \beta}, \quad Z = \hat{C} + \frac{1}{\alpha}(\frac{C-W}{W})$
   - **[WPCMS]** $\Theta := \text{PCMS}_W(Z, H)$

PoisMS examples

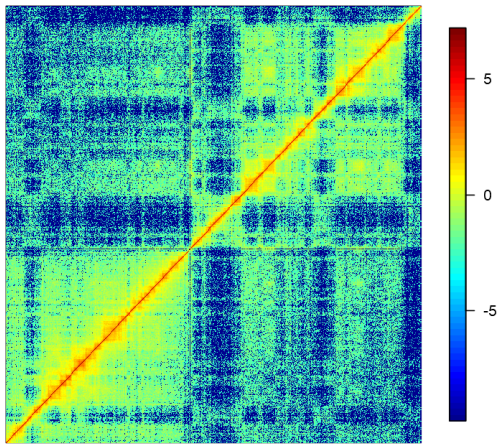
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$

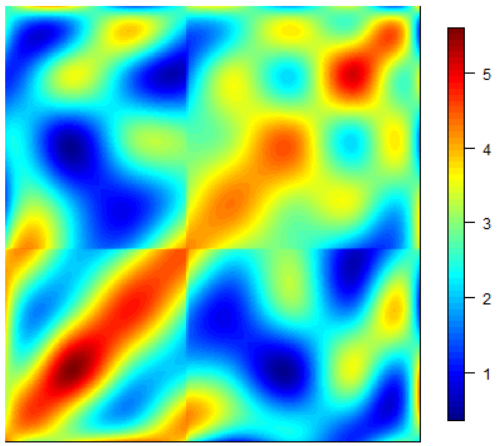
PoisMS examples ($df = 10$)

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

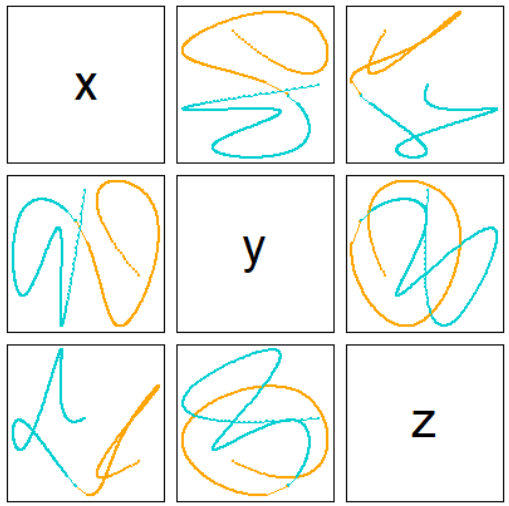
PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$



Original data



Reconstruction

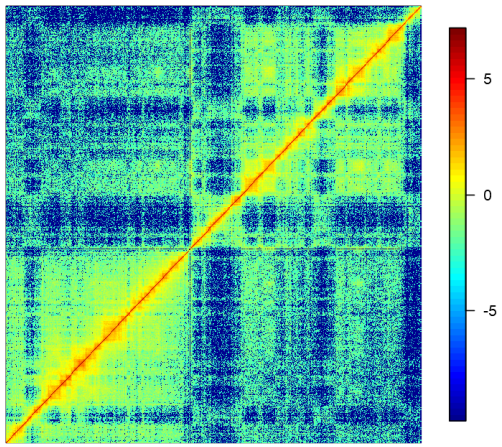


3D conformation

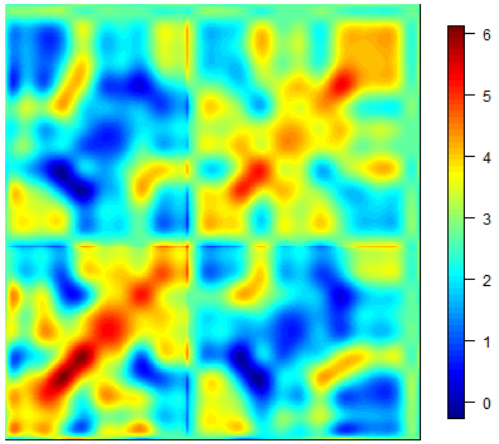
PoisMS examples ($df = 25$)

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

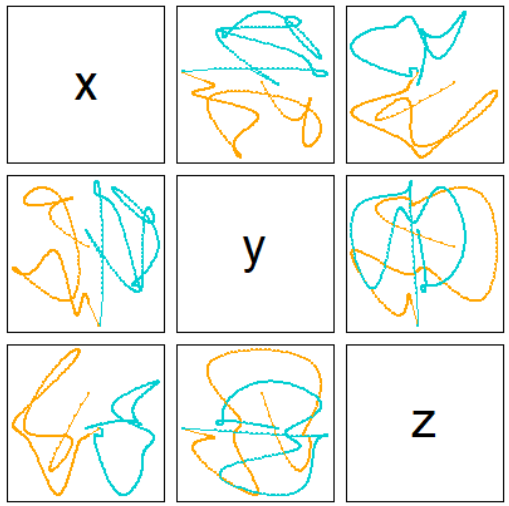
PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$



Original data



Reconstruction

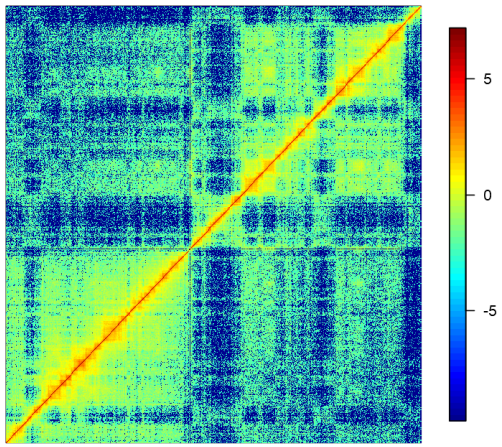


3D conformation

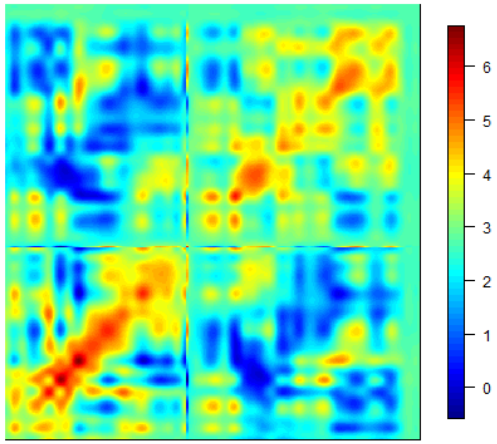
PoisMS examples ($df = 50$)

Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus, chromosome 20, probe resolution 100kb, $n = 625$.

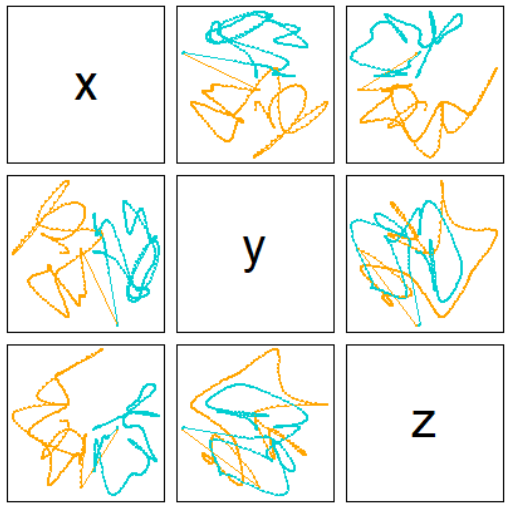
PoisMS parameters: $\alpha = 1$ and $\beta = \log\left(\frac{\sum_{i,j=1}^{n} C_{ij}}{n}\right)$



Original data



Reconstruction



3D conformation

References

- T. Stevens et al. *3D structures of individual mammalian genomes studied by single-cell Hi-C*. Nature, 544:59–64, 2017.
- N. Varoquaux et al. *A statistical approach for inferring the three-dimensional structure of the genome*. Bioinformaics, 30(12):26-33, 2014.
- Z. Zhang et al. *3D Chromosome Modeling with Semi-Definite Programming and Hi-C Data*. Journal of computational biology, 20(11):831–846, 2013.
- C. Zou et al. *HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure*. Genome Biology, 17(40), 2016.
- M. Rosenthal et al. *Bayesian Estimation of 3D Chromosomal Structure from Single Cell Hi-C Data*. Journal of Computational Biology, published online, 2019.
- A. Buja et al. *Data Visualization With Multidimensional Scaling*. Journal of Computational and Graphical Statistics, 17(2):444-472, 2008.
- R. Mazumder et al. *Spectral Regularization Algorithms for Learning Large Incomplete Matrices*. Journal of Machine Learning Research, 11: 2287-2322, 2010.
- N. Srebro et al. *Weighted Low-Rank Approximations*. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

Thank you for your attention!