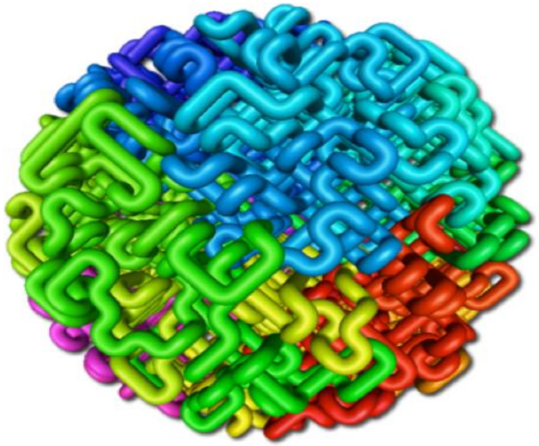
Statistical Topology of Genome Analysis: From Chromosome Conformation Capture data to 3D structure



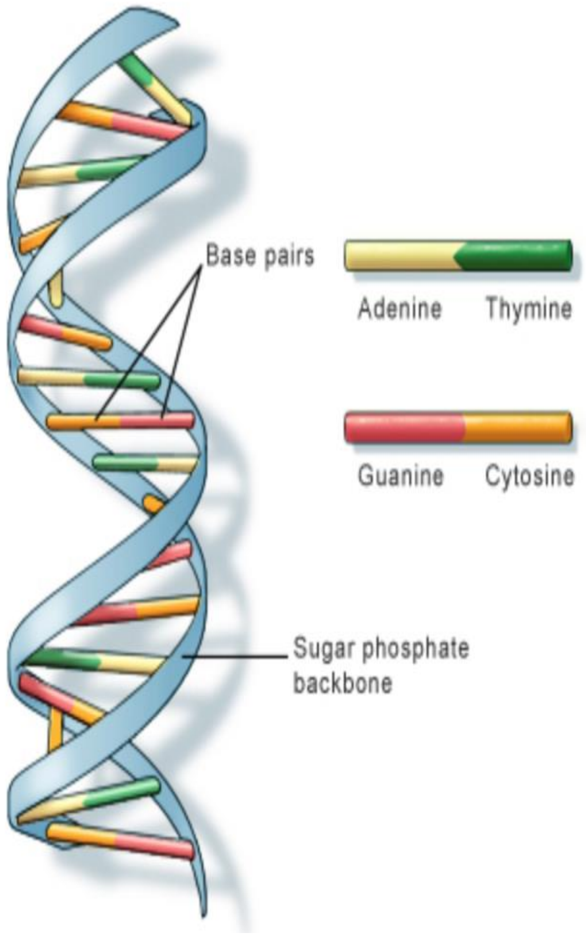
Maxime Pouokam

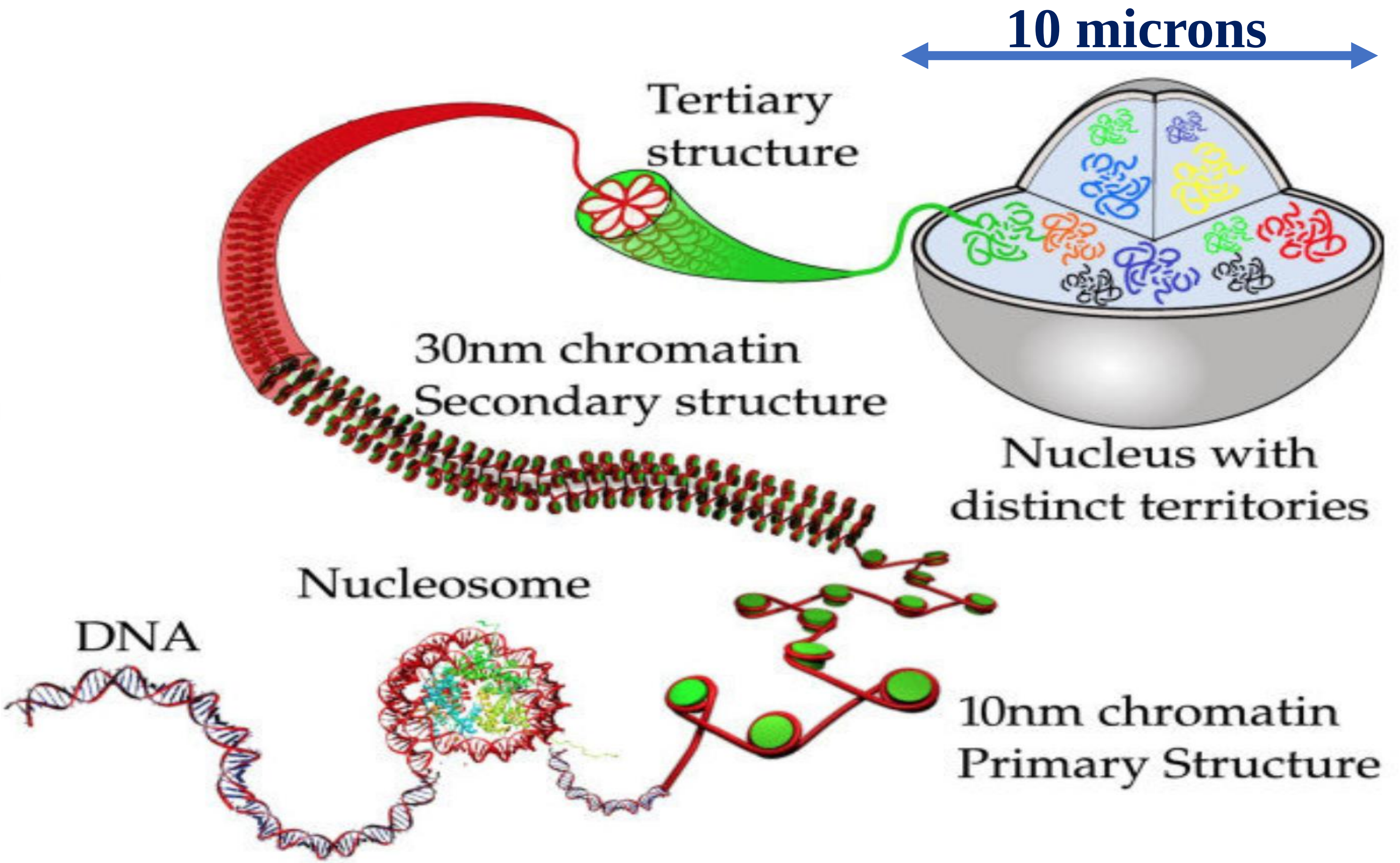
UC Davis

Statistics department

The genome folding problem

How can a 2 meters of DNA being packed into a 10 um diameter cell ?

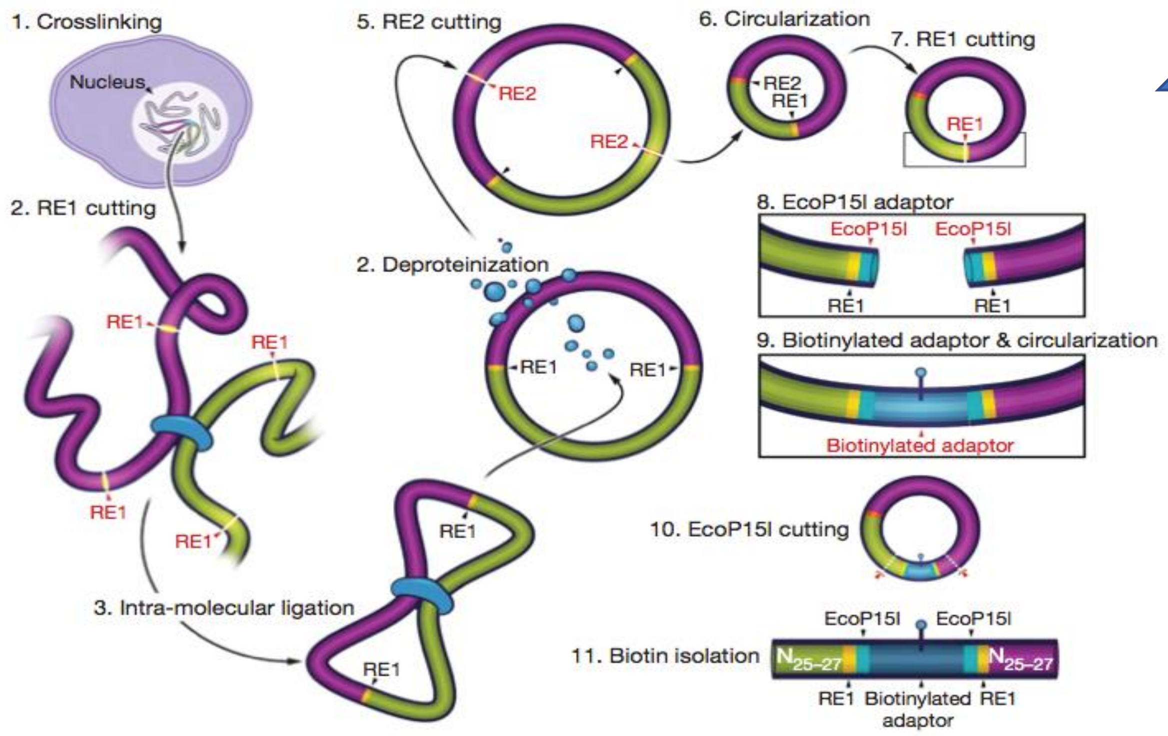


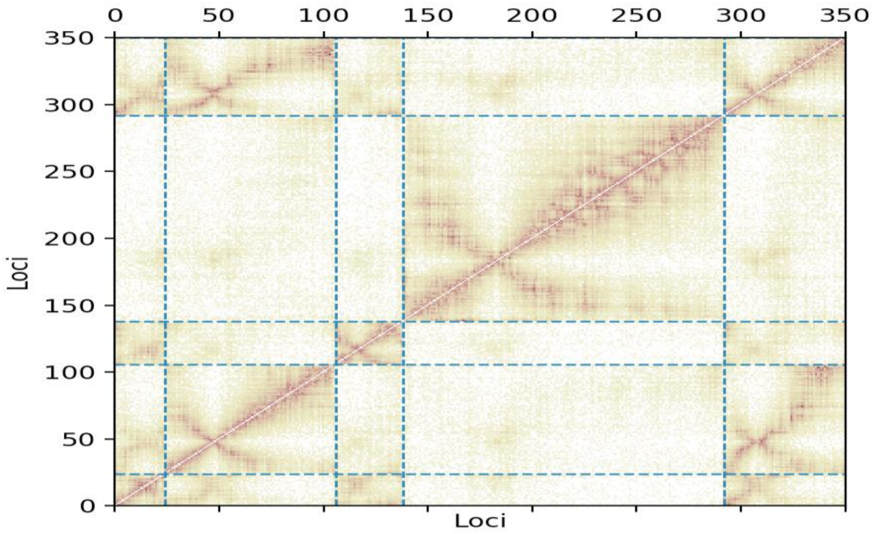


Credit: U.S. National Library of Medicine

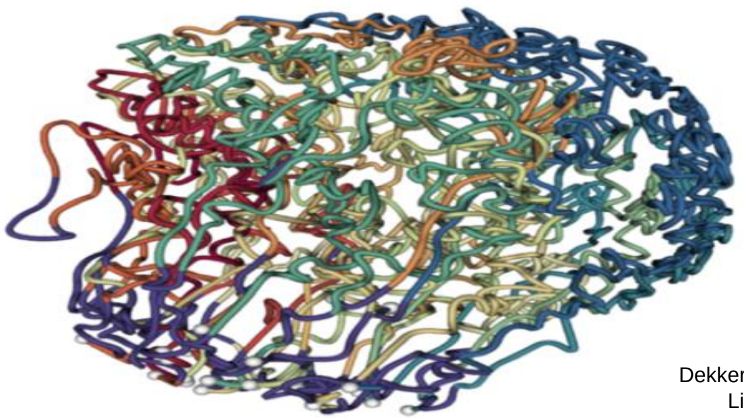
Chromosome Conformation Capture-based assays measure proximal pairs of DNA loci

3C / $\boxed{\text{4C}}$ / 5C / Hi-C / TCC





Multidimensional scaling (MDS)



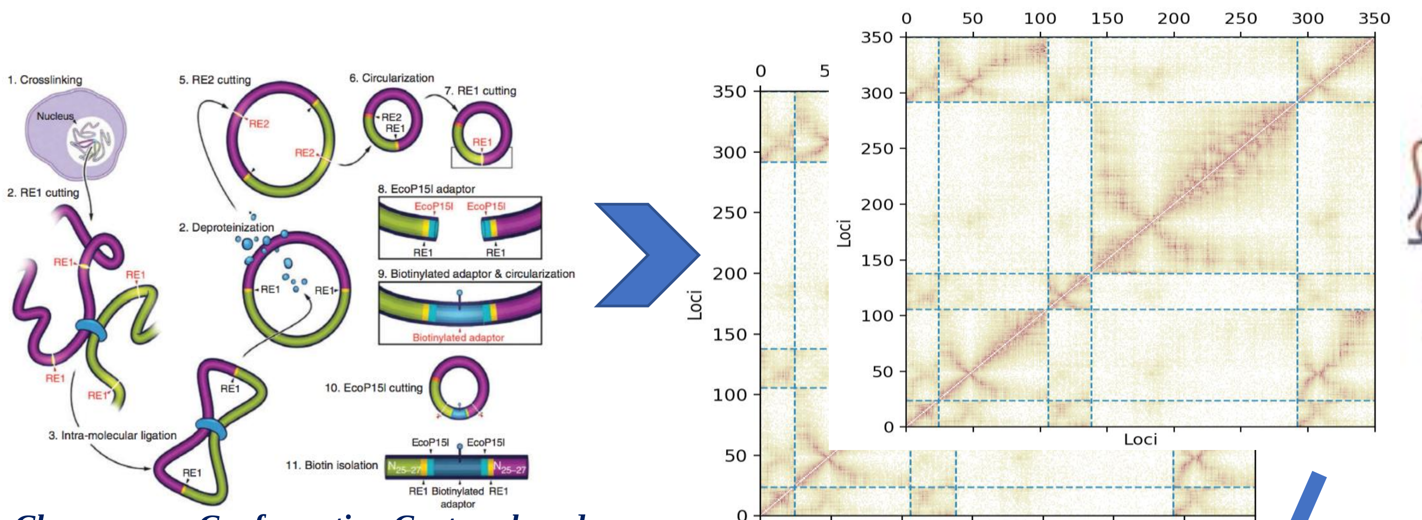
3D Yeast S.Cerevisiae

Dekker et al. *Science* 2002, Lieberman-Aiden, Van Berkum et al.,Science 2009 Rao et al., Cell 2014

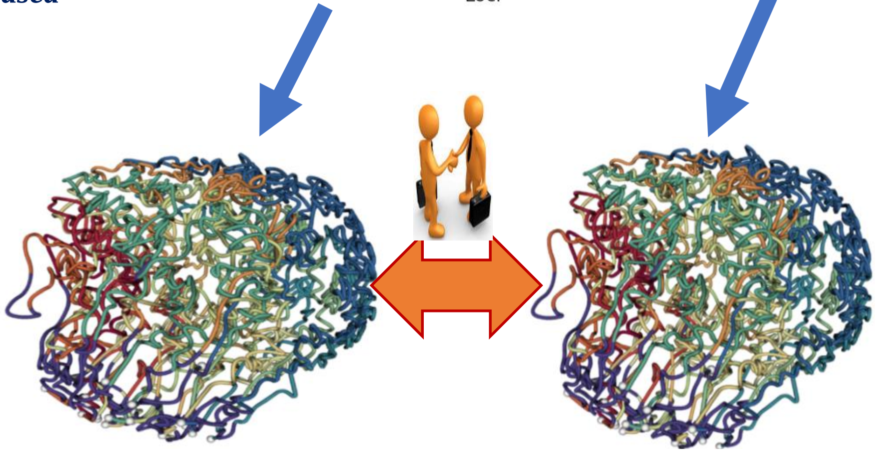
Data: 1-10 billion sequencing reads

- ***Technical bias (from the sequencing and mapping)***
- ***Biological bias (inherent to the physical properties of chromatin)***

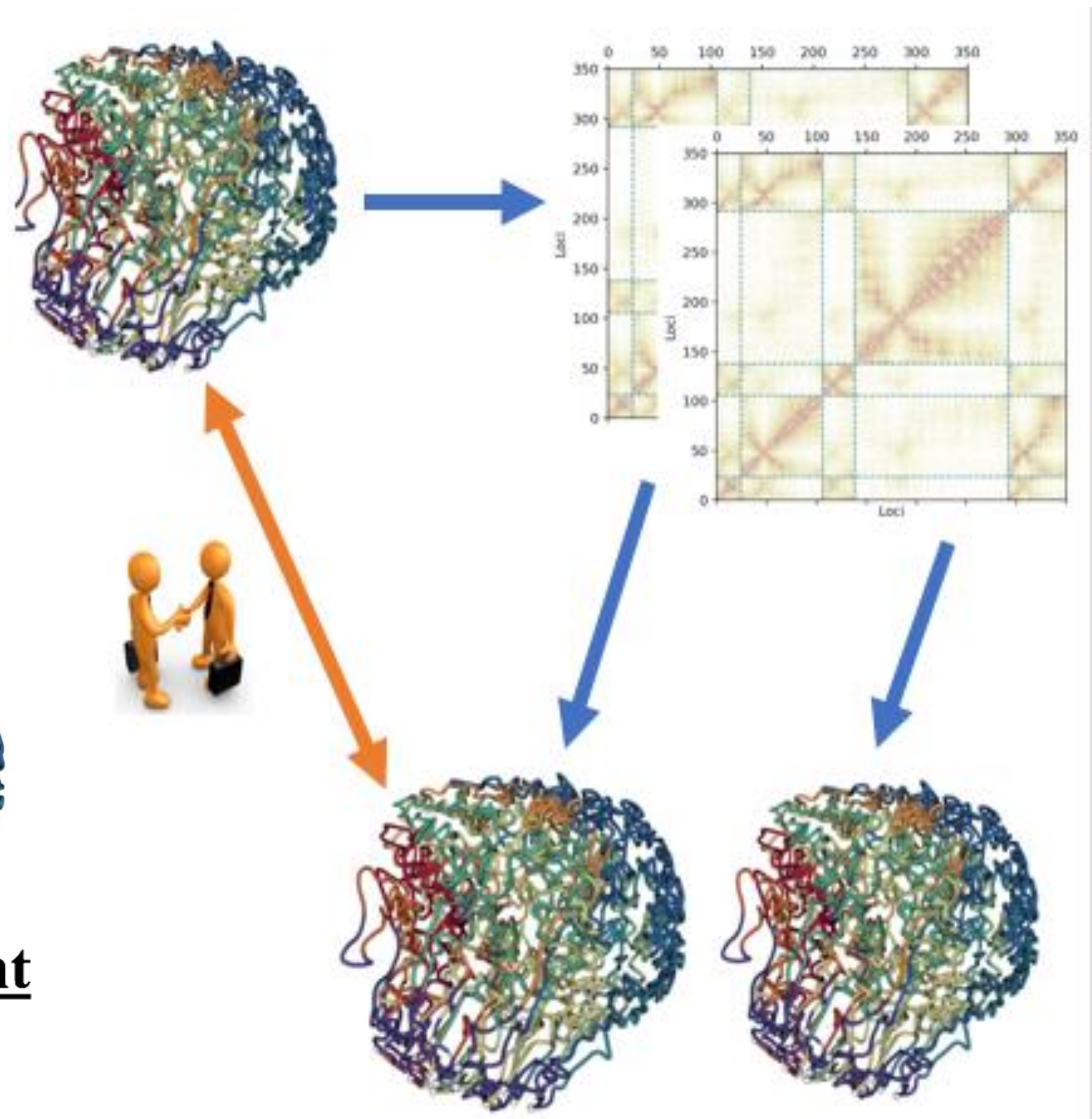
Reproducibility / Evaluation problem



Chromosome Conformation Capture-based



3D S. cerevisae reconstructions



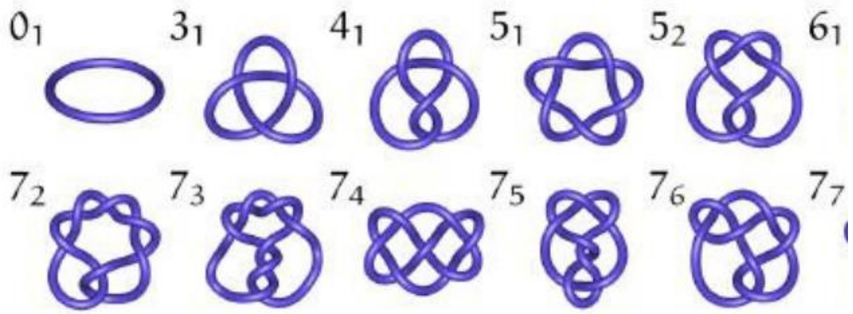
Two components needed for agreement assessment

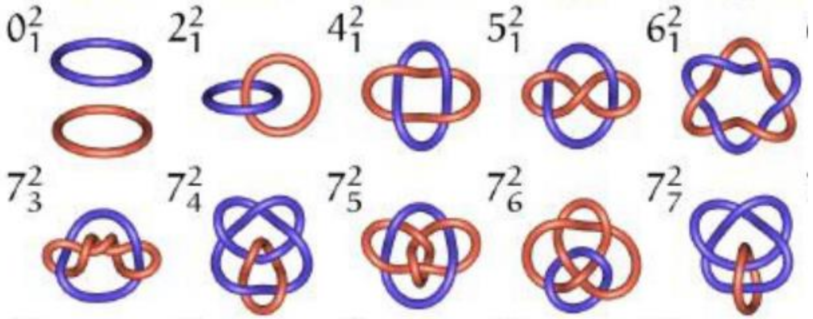
- **Metric that measures closeness**
- **Null referent distribution for statistical significance**

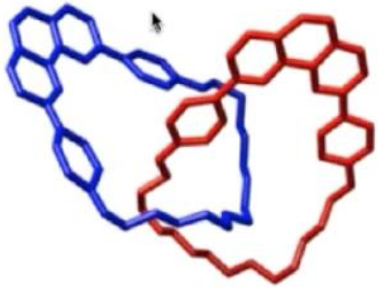
Knot / Link exists in nature

NATURE

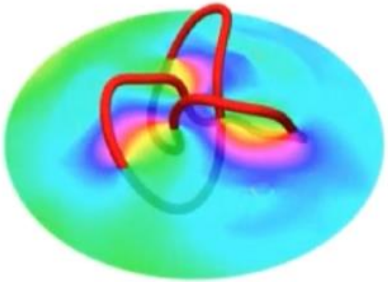
ART



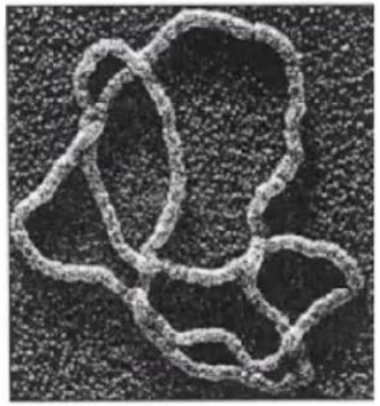




CHEMISTRY

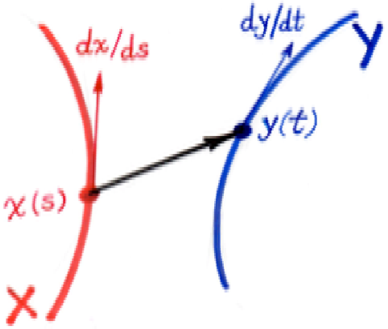


PHYSICS



BIOLOGY

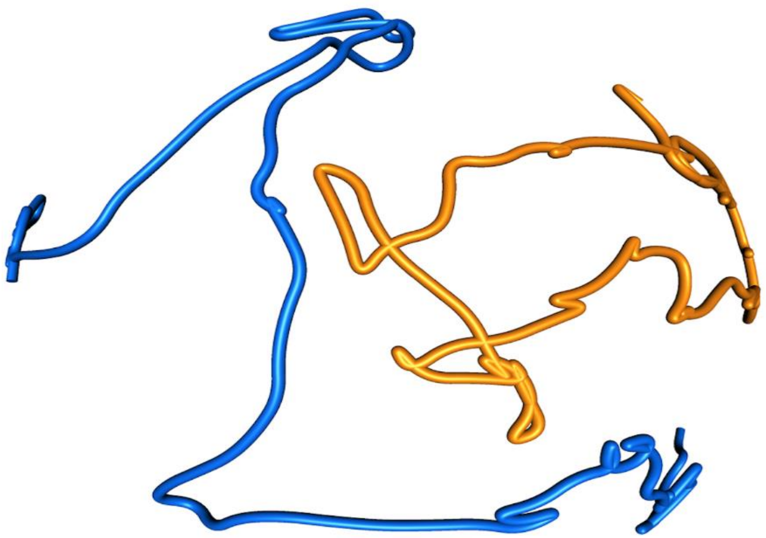
Linking Number is a topological Invariant – use to compute entanglements

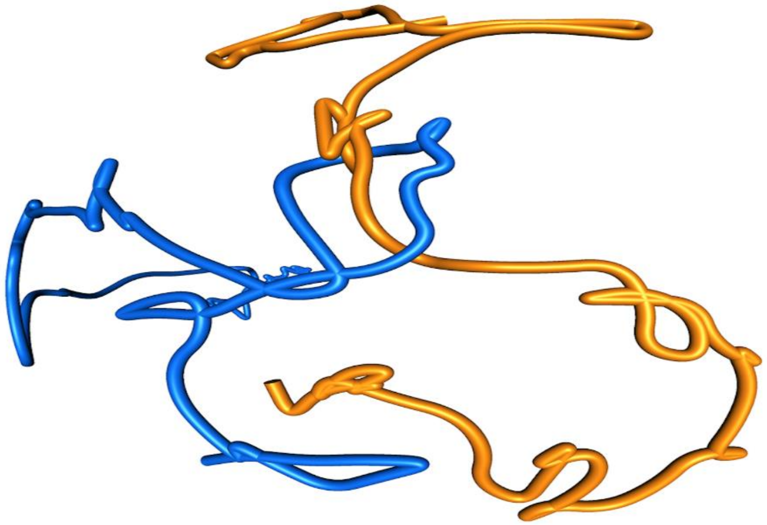


Gauss double integral of the two curves is defined as

$$Lk = \frac{1}{4\pi}\int_X \int_Y \frac{x(s) - y(t)}{|x(s) - y(t)|^3} dx(s) \times dy(t)$$

It is difficult to measure entanglement in open chains





Topologically can be deformed into

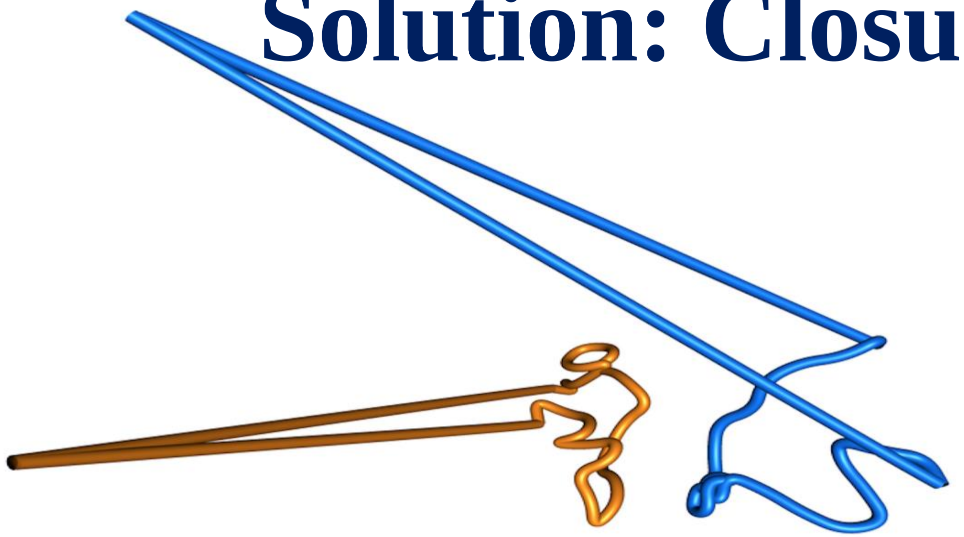
<u>Desired properties</u>

- *Computable, Well defined, Interpretable*
- *Stable: minimum effects from small perturbations*
- *Be continuous in "some sense"*

Pouokam et al. 2019 **Pouokam et al. (in prep)**

Refs: E. Rawdon, K. Millett, J. Sulkowska …

Solution: Closure



Closure algorithm: For a given pair of chromosomes i and j, X_{ij} represents a random outcome of determining the topological state of the two chromosomes,

$$X_{ij} = \begin{cases} 1, & \text{if the ith and jth circularized chromosomes have non zero } Lk \\ 0, & \text{otherwise} \end{cases}$$

Define $Y_{ij} = \sum_{n=1}^{N} (X_{ij})_n$ is the total number of times the LK of the two circularized chromosomes i and j were found to be nonzero.

p_{ij} is the linking proportion associated to chromosomes i and j and estimated as $\hat{p}_{ij} = \frac{Y_{ij}}{N}$

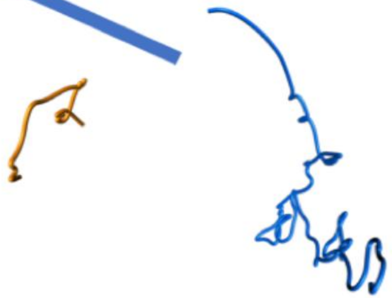
Results: The linking proportions (Lp) measure entanglement between pair of chromosomes and are lower than expected…

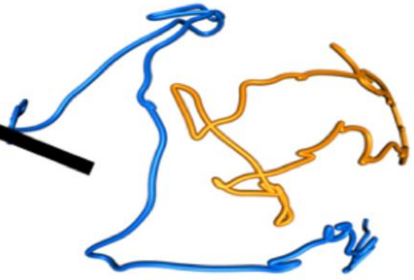
Lp are recorded in %.

Lp > 50% in red

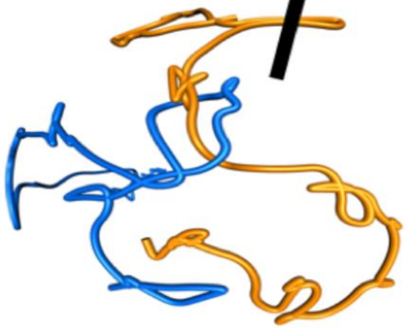
Reconstruction 8

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI
I	-	6.2	2.9	9.4	11.8	2.5	19.8	8.7	2.3	3.9	6.2	6.1	3.7	7.0	6.1	6.9
II	7.6	-	3.3	30.3	7.5	14.5	9.4	9.6	9.9	6.0	7.7	14.0	3.8	10.5	11.1	60.8
III	2.3	5.7	-	4.5	3.7	2.0	4.3	2.8	3.3	3.3	3.0	5.8	6.7	3.7	7.8	2.7
IV	7.1	17.3	7.0	-	13.1	11.0	19.7	44.7	12.7	6.2	13.2	18.6	5.3	15.0	22.3	41.8
V	6.1	11.1	2.8	13.9	-	3.4	17.3	11.7	4.4	7.7	7.7	11.6	5.4	16.1	11.4	9.5
VI	1.5	2.6	2.9	8.5	2.2	-	5.7	4.2	3.8	2.6	3.3	6.1	2.5	4.6	6.9	12.5
VII	11.8	14.8	2.6	11.9	12.2	2.4	-	15.1	5.8	7.4	8.5	11.7	5.0	16.7	10.2	13.4
VIII	11.2	13.3	2.7	10.2	7.2	1.7	51.8	-	5.0	5.9	11.9	11.3	2.1	19.8	7.5	10.9
IX	4.4	5.0	2.1	5.7	2.8	1.5	3.5	3.4	-	6.4	4.9	21.5	4.7	5.7	9.7	11.0
X	5.6	12.3	2.5	12.6	30.8	1.6	12.5	11.9	3.3	-	7.0	34.8	10.4	13.0	13.7	7.0
XI	31.5	16.1	3.0	11.2	9.4	1.7	18.2	17.5	4.5	10.5	-	17.6	3.4	18.6	15.6	16.1
XII	23.0	24.8	5.6	55.9	16.4	4.5	33.3	70.3	9.7	17.6	35.4	-	8.0	19.5	35.8	26.4
XIII	6.0	24.9	4.5	16.0	9.9	2.7	11.6	9.7	3.7	8.9	9.4	20.2	-	5.5	5.6	3.8
XIV	9.8	11.3	2.6	13.6	10.2	3.5	24.7	9.5	12.2	8.6	12.7	29.8	8.9	-	19.3	10.2
XV	9.5	21.9	4.3	13.3	12.2	3.7	50.9	20.9	4.9	15.8	15.7	39.1	99.4	14.7	-	23.7
XVI	6.9	8.9	2.2	12.1	16.4	2.9	15.5	9.8	3.0	12.7	8.1	16.1	6.7	13.2	13.9	-

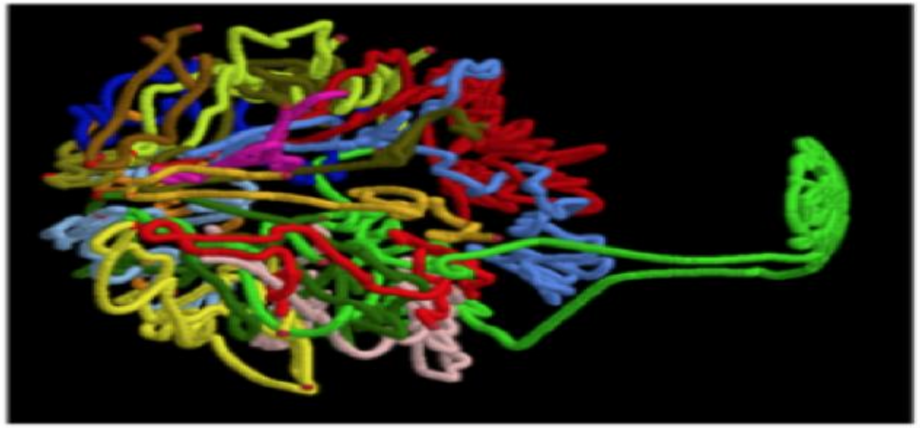


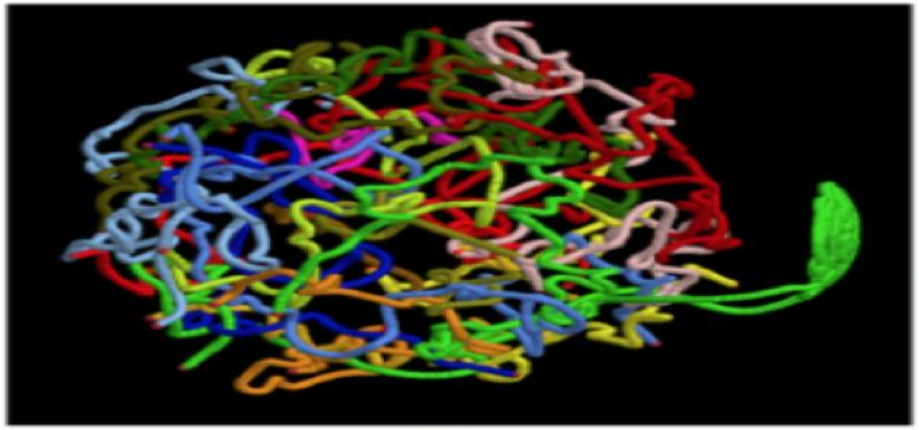


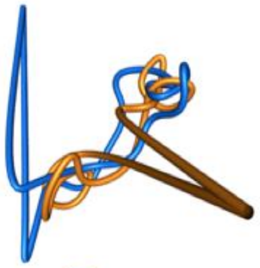
Reconstruction 7



Our statistical agreement approach

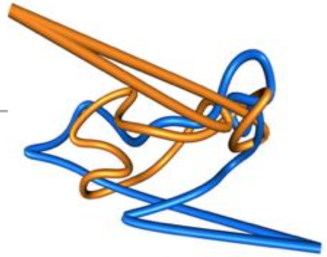






	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	5.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	4.9	9.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	5.8	41.8	8.1	-	-	-	-	-	-	-	-	-	-	-	-	-
5	6.3	37.8	7.2	19.1	-	-	-	-	-	-	-	-	-	-	-	-
6	7.3	53.8	17.2	92.8	16.5	-	-	-	-	-	-	-	-	-	-	-
7	21.6	18.8	13.1	20.1	15.7	19.3	-	-	-	-	-	-	-	-	-	-
8	6.4	21.5	6.6	42.6	26	9.4	23	-	-	-	-	-	-	-	-	-
9	4.7	20.9	11.3	10.4	23.7	14.3	17.7	12.7	-	-	-	-	-	-	-	-
10	8.9	14.5	7	18.9	13.8	9.8	49.4	20.3	14.3	-	-	-	-	-	-	-
11	25.1	11.7	10.4	15.3	15.6	10.6	27.9	17.3	44.4	41.1	-	-	-	-	-	-
12	19.9	31.8	11.2	30.7	17.4	21.5	48.1	55.4	17.2	21.9	18.6	-	-	-	-	-
13	9.6	26.3	10	92.1	22.2	21.4	30.2	37.7	22.6	32.1	33.2	75.2	-	-	-	-
14	9.2	19.2	8.3	56.7	44.7	10.6	26.5	53.8	18.9	34.9	29.1	47	70.1	-	-	-
15	9.6	31.2	44.5	24.9	35.1	40.8	27.5	25	53.2	16.9	19.7	35.3	25	23	-	-
16	13.2	19.4	12.7	19.2	16.5	20.4	56.8	19.5	16	60.5	76.5	35.9	23.6	45.9	26.2	

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	7.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	1.9	8.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	5.7	25.0	9.6	-	-	-	-	-	-	-	-	-	-	-	-	-
5	4.9	13.2	3.2	17.5	-	-	-	-	-	-	-	-	-	-	-	-
6	1.4	2.7	2.8	8.2	2.1	-	-	-	-	-	-	-	-	-	-	-
7	15.7	16.5	3.6	14.6	14.6	3.7	-	-	-	-	-	-	-	-	-	-
8	15.4	13.5	4.2	11.4	10.4	2.0	48.7	-	-	-	-	-	-	-	-	-
9	7.2	5.1	1.6	6.4	3.4	1.8	4.5	3.0	-	-	-	-	-	-	-	-
10	8.5	10.2	3.8	12.9	37.9	2.1	15.9	15.9	4.3	-	-	-	-	-	-	-
11	40.2	18.5	3.3	11.7	8.8	2.3	22.7	20.9	5.8	10.3	-	-	-	-	-	-
12	24.4	25.1	6.8	43.6	23.4	4.6	35.0	46.9	11.2	20.1	55.6	-	-	-	-	-
13	7.4	29.9	6.4	16.9	11.6	2.0	11.9	10.6	4.1	11.2	10.9	18.9	-	-	-	-
14	12.4	13.4	3.3	17.5	10.5	2.8	30.0	10.8	17.8	9.7	14.4	32.7	9.6	-	-	-
15	10.0	26.9	4.8	20.0	17.6	4.0	45.1	24.0	4.5	16.4	19.0	58.8	99.3	11.5	-	-
16	7.3	9.9	3.1	12.6	19.2	4.4	19.0	9.8	3.2	14.9	10.7	17.6	9.0	17.6	17.5	-



Statistic / probability measure metric / distance

test to assign p-values: correct p-values for multiple testing

Model and test formulation

Model,

$$Y_{ij}^k \sim Bin(N, p_{ij}^k); \qquad 1 \le i < j \le 16;$$

Hypothesis testing,

$$H_0 : p_{ij}^k = p_{ij}^l \qquad VS. \qquad H_1 : p_{ij}^k \neq p_{ij}^l; \qquad k \neq l.$$

The Likelihood ratio test (LRT) is defined as,

$$\lambda(Y) = \frac{sup\{L(\theta; Y) : \theta \in \Theta_0\}}{sup\{L(\theta; Y) : \theta \in \Theta\}}, \qquad Y = (Y_{ij}^k, Y_{ij}^l), \qquad \theta = (p_{ij}^k, p_{ij}^l)$$

Thus the LRT statistics is, $\lambda(Y) = \dfrac{L(\hat{\theta}_0; Y)}{L(\hat{\theta}; Y)}$

By Wilks' Theorem(1938), under H_0, $-2log(\lambda(Y)) \xrightarrow{D} \chi^2_{120}$

Pearson Chi-square test statistic

For a pair of reconstructions k and l, the Pearson Chi-square test statistic is,

$$\mathbf{X}^2 = \sum_{l=1}^{2} \sum_{i<j} \frac{(O_{ijl}^{k} - E_{ijl})^2}{E_{ijl}} + \sum_{l=1}^{2} \sum_{i<j} \frac{(O_{ijl}^{l} - E_{ijl})^2}{E_{ijl}}$$

- O_{ij1}^{k} = observed number of linked conformations out of N in the closure algorithm (which is Y_{ij}^{k} in our notation)
- O_{ij2}^{k} = number of unlinked conformations out of N, $N - Y_{ij}^{k}$
- E_{ij1} = expected number of linked conformations out of N, $E_{ij1} = \dfrac{Y_{ij}^{k} + Y_{ij}^{l}}{2}$
- E_{ij2} = expected number of unlinked conformations, $E_{ij2} = N - \dfrac{Y_{ij}^{k} + Y_{ij}^{l}}{2}$

Hence, $\mathbf{X}^2 = 2N \sum_{i<j} \dfrac{(Y_{ij}^{k} - Y_{ij}^{l})^2}{[Y_{ij}^{k} + Y_{ij}^{l}][2N - (Y_{ij}^{k} + Y_{ij}^{l})]}$ Under H_0, $\mathbf{X}^2 \xrightarrow{D} \chi_{120}^2$

Conclusion

The Likelihood Ratio test and Pearson Chi-Square test separated all reconstructions obtained by MDS methods

Most p-values << 0.0001

Semi-soft thresholding approach for inference of proportions

- We define, $\delta_{ij} = p_{ij} - q_{ij}$, $\quad z_{ij} = \dfrac{\hat{\delta}_{ij}}{\sqrt{\dfrac{\hat{p}_{ij}(1-\hat{p}_{ij}) + \hat{q}_{ij}(1-\hat{q}_{ij})}{N}}} \quad i < j$
- The shrinkage variable as, $\tilde{\delta}_{ij}(c) = \hat{\delta}_{ij} G(|z_{ij}|/c)$
- The squared error distance, $F(c) = \sum_{i<j} (\tilde{\delta}_{ij}(c) - \delta_{ij})^2$

$$F(c) = \sum_{i<j} \tilde{\delta}^2_{ij}(c) + \sum_{i<j} \delta^2_{ij} - 2 \sum_{i<j} \tilde{\delta}_{ij}(c) \delta_{ij}$$

The criterion function is formulated as,

$$\hat{F}(c) = \sum_{i<j} \tilde{\delta}^2_{ij}(c) - 2 \sum_{i<j} \hat{\delta}_{ij} \tilde{\delta}_{ij}(c) + 2 \sum_{i<j} \{ \hat{Var}(\hat{p}_{ij}) \frac{\partial \tilde{\delta}_{ij}(c)}{\partial \hat{p}_{ij}} - \hat{Var}(\hat{q}_{ij}) \frac{\partial \tilde{\delta}_{ij}(c)}{\partial \hat{q}_{ij}} \}$$

We use known distribution functions G_1 and G_2 on $[0, \infty)$ where,
$G_1(u) = u^2/(1+u^2)$, and $G_2(u) = (u-0.5)^2_+/[1+(u-0.5)^2_+]$

Semi-soft thresholding approach discriminates all MDS reconstructions

Table 3.21: Number of zero entries in the vectors $\hat{\delta}$ and its shrinkage analogues, $\tilde{\delta}(\hat{c})$ obtained using the CLT and Arsine transformation. The smoothing function is $G_1 = u^2/(1+u^2), u > 0$.

		CLT	Arcsine
Reconstructions	**% of zeros in $\hat{\delta}$**	**% of zeros in $\tilde{\delta}(\hat{c})$**	**% of zeros in $\tilde{\delta}(\hat{c})$**
1 & 2	8.33	15.00	14.17
1 & 3	5.83	5.83	12.50
1 & 4	10.83	14.17	13.33

Table 3.22: Number zero of entries in the vectors $\hat{\delta}$ and its shrinkage analogues, $\tilde{\delta}(\hat{c})$ obtained using the CLT and Arsine transformation. The smoothing function is $G_2 = (u-0.5)_+^2/[1+(u-0.5)_+^2], u > 0$.

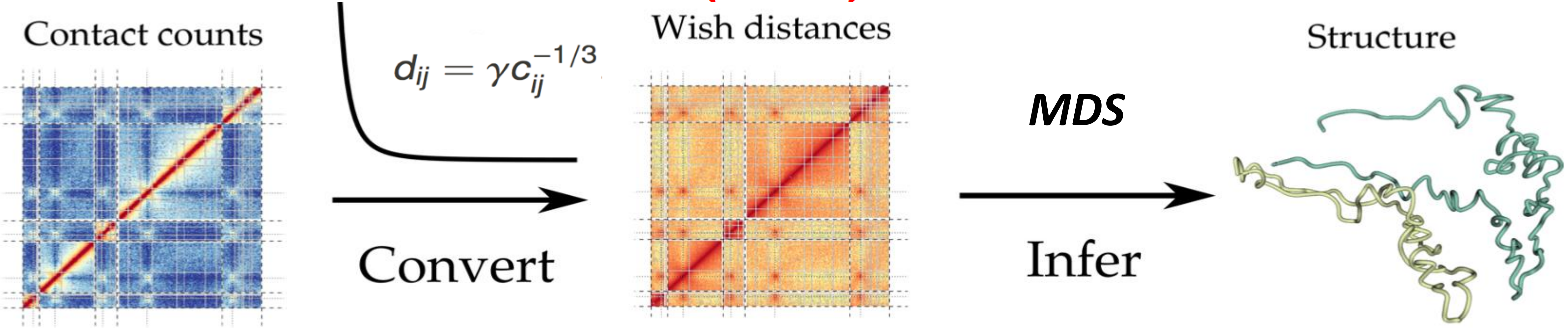
		CLT	Arcsine
Reconstructions	**% of zeros in $\hat{\delta}$**	**% of zeros in $\tilde{\delta}(\hat{c})$**	**% of zeros in $\tilde{\delta}(\hat{c})$**
1 & 2	8.33	15.83	16.67
1 & 3	5.83	5.83	15.00
1 & 4	10.83	14.17	13.33

Conclusion

MDS-based reconstruction approaches fail to preserve chromosomal topology

3D reconstruction as dimension reduction, $O(N^2)$ to $O(3N)$, with N being the number of genomic loci/beads

Metric multidimensional scale (MDS)



Formulation

$$\underset{\mathbf{x}_1,\ldots,\mathbf{x}_n}{\text{minimize}} \quad \sigma(\mathbf{X}, C) = \sum_{i,j|c_{ij}\neq 0} w_{ij}(\|x_i - x_j\|_2 - \Theta(c_{ij}))^2$$

$$\text{s.t.} \quad \text{some constraints}$$

- $\mathbf{X}$: 3D coordinates
- $\mathbf{C}$: normalized contact counts.

CCC data

Table 1.1: Frequency of interactions within chromosome I (top part) and between chromosome I and chromosome II (bottom part) using 4C.

Chromosome	Locus 1	Chromosome	Locus 2	Contact frequency	Q-value
I	2894	I	191604	8	7.964353e-03
I	2894	I	226931	11	2.141016e-05
I	3437	I	31834	47	8.402414e-04
I	3437	I	167621	10	8.193970e-04
I	3437	I	226931	9	7.598729e-04
I	5091	I	26147	174	2.123039e-39
⋮	⋮	⋮	⋮	⋮	⋮

Smooth3D, Inferring 3D structure of genome via cubic spline approximation

Model,

$$Y_{ij} = \log(c_{ij}) = \log(\mu_{ij}) + \varepsilon_{ij}, \; j = 1, \ldots, n_i, \;\; i = 1, \ldots, k,$$

where $\{\varepsilon_{ij}\}$ are i.i.d. with mean zero and variance σ^2.
Our goal is to find a 3D curve $\boldsymbol{x}$ from $[0, 1]$ to $\mathbb{R}^3$ so that

$$Q(\boldsymbol{x}) = \sum_{i,j} [Y_{ij} + \alpha \log ||\boldsymbol{x}(t_i) - \boldsymbol{x}(t_j)||]^2$$

is minimized. Where $\mu_{ij} = ||\boldsymbol{x}(t_i) - \boldsymbol{x}(t_j)||^{-\alpha}$, $\;\; t_i$ is the position of locu i.

$$\begin{gathered} x_1(0) = x_2(0) = x_3(0) = 0, \\ x_1(0.5) = 0, x_1(1) = x_2(1) = 0, \\ \int x_1(t) \leq -\delta, \quad x_2(0.5) \leq -\delta, \; x_3(1) \geq \delta, \end{gathered} \tag{1}$$

where $\delta > 0$ is a very small real number.

Spline Parametrization of $x(t)$

We use cubic B splines to model the curve $\boldsymbol{x}$

$$x_1(t) = \boldsymbol{\beta}_1^T \boldsymbol{B}_1(t),\ x_2(t) = \boldsymbol{\beta}_2^T \boldsymbol{B}_2(t),\ x_3(t) = \boldsymbol{\beta}_3^T \boldsymbol{B}_3(t), \tag{2}$$

where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ are $\boldsymbol{\beta}_3$ are $k+1, k+2$ and $k+3$ dimensional vectors respectively. k **is the # of knots**. The inequality constraints in (1) are now

$$\sum \beta_{1j} \leq -\delta,\ \boldsymbol{\beta}_2^T \boldsymbol{B}_2(0.5) \leq -\delta,\ \boldsymbol{\beta}_3^T \boldsymbol{B}_3(1) \geq \delta. \tag{3}$$

Thus the optimization problem is to minimize

$$Q(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \alpha) = \sum_{i,j} [Y_{ij} + \alpha \log ||\boldsymbol{x}(t_i) - \boldsymbol{x}(t_j)||]^2\,, \tag{4}$$

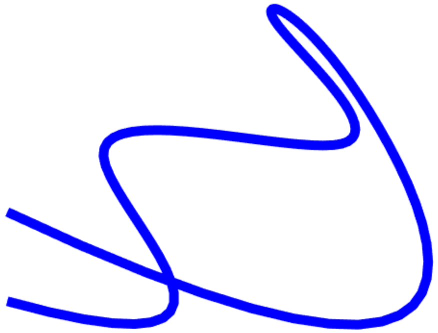
with respect to $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ and α, subject to the constraints given in (3).

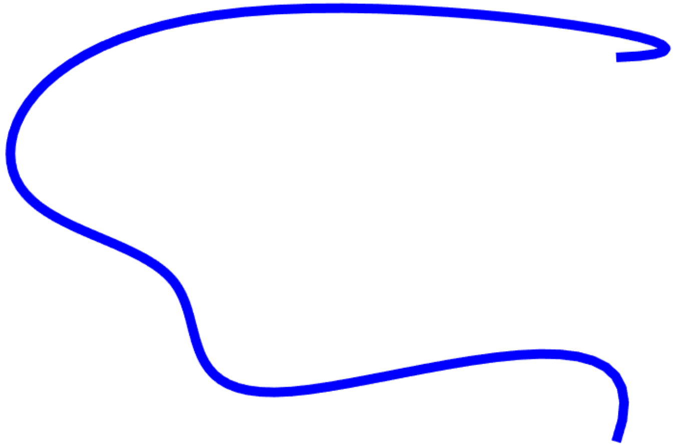
Global minimum: We used the **random multistart method**. In the outer loop we obtained random starting points. For each starting point, in the inner loop, we use **cyclic block-coordinate minimization** in order to obtain a **local minima**.

Table 4.1: Dimension reduction of Smooth3D as compared to the MDS-based and other optimization-based approaches.

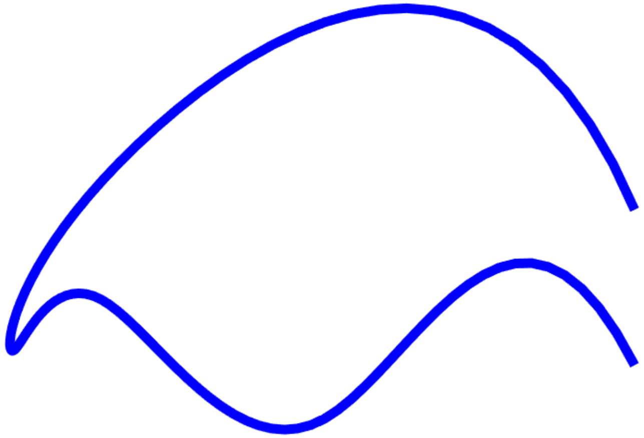
	MDS-based methods				**Smooth3D**	
	# of loci at the loci resolution	# of parameters at the loci resolution	# of beads at a 10 kb resolution	# of parameters 10 kb resolution	# of knots	# of parameters at the loci resolution
Chrom I	47	$47 \times 3 = \mathbf{141}$	23	$23 \times 3 = \mathbf{69}$	$k = 4$	$3 \times 4 + 6 = \mathbf{18}$
Chrom II	239	$239 \times 3 = \mathbf{717}$	80	$80 \times 3 = \mathbf{240}$	$k = 5$	$3 \times 5 + 6 = \mathbf{21}$

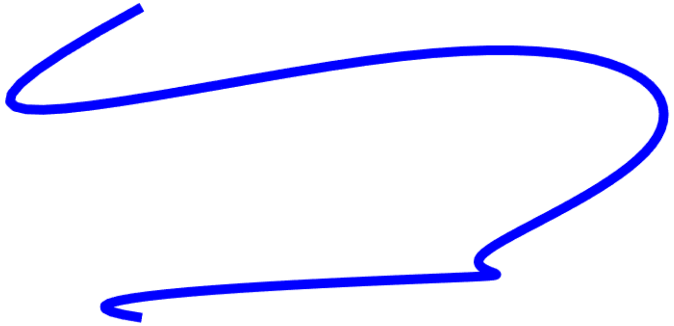
Result: Differing views of 3D Reconstruction structure of chromosome I



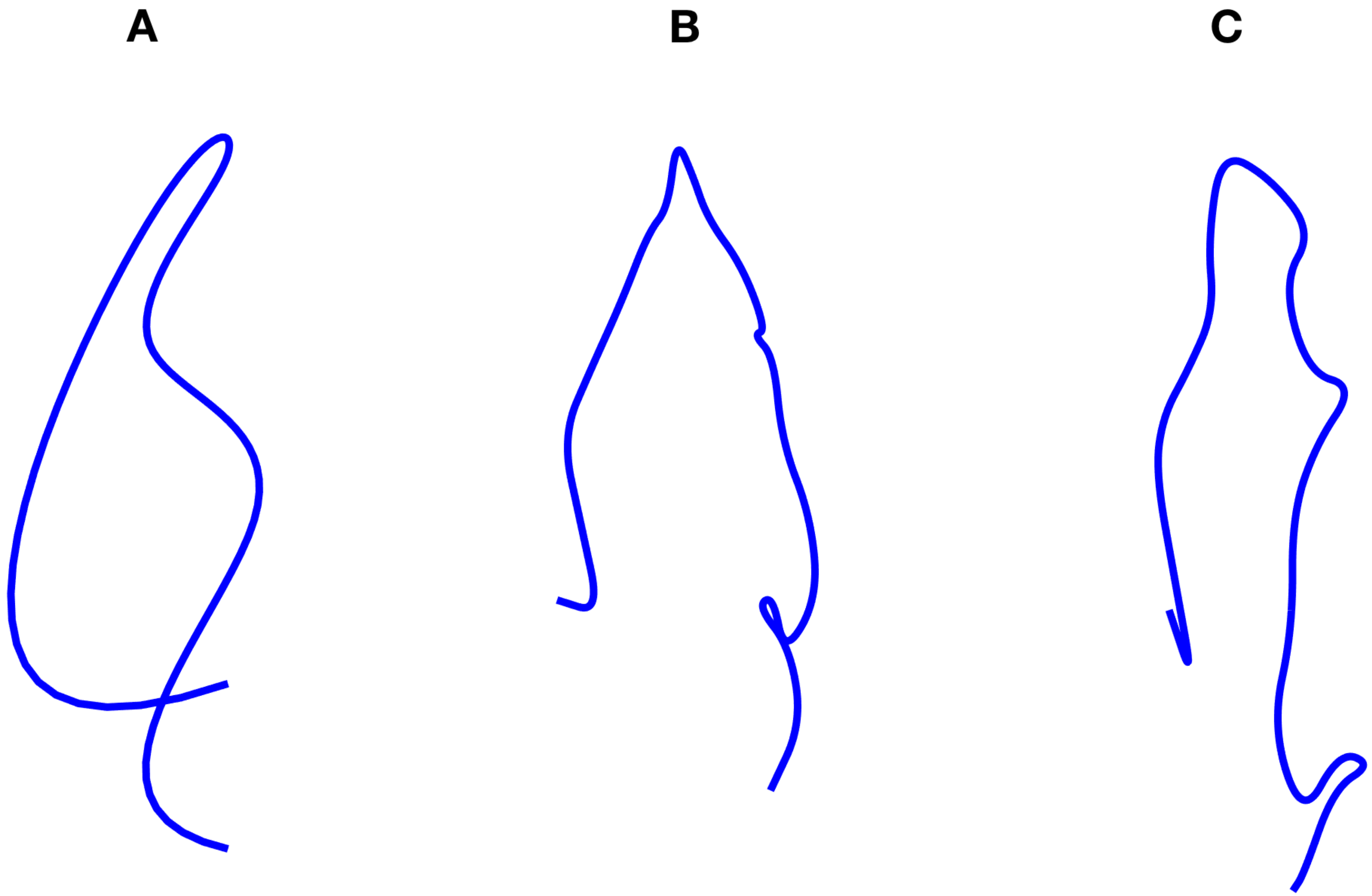


- $\alpha_0 = 3, \quad \hat{\alpha} = 1.015$
- The minimum value was $Q = 48.841$





Chromosome I from Smooth3D (A) and MDS methods (B, C)



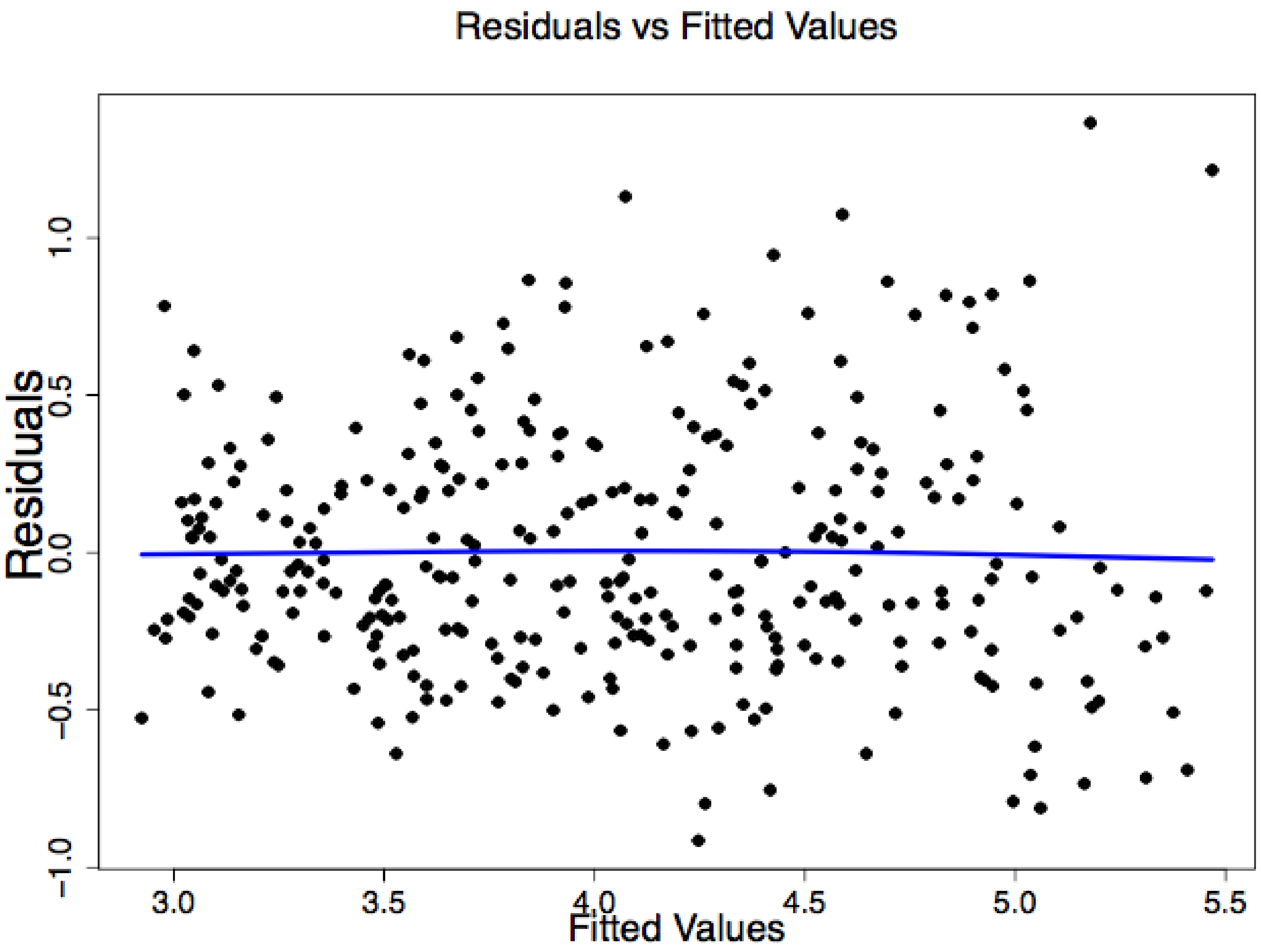


Figure 4.2: Residuals versus fitted values for Chromosome I

Conclusion

- Obtaining 3D genome conformation is important
- Reconstruction methods are challenging
- Many 3D reconstructions are consistent with any given contact map (**Optimization methods – local minimal**)
- Can be diagnosed by **comparing obtained solution** under perturbed data inputs, constraint specifications, starting conditions
- **Measuring entanglement** can help exploring topological state of genome reconstructions
- In **agreement assessment of 3D reconstructions,** the **metric** is as important as the **referent distribution**
- **Before any downstream functional analysis could be made we need reconstruction methods should be fast and stable, Smooth3D ???!!!**

Acknowledgements





Prof. Javier Arsuaga, co-advisor

Prof. Burman Prabir, co-advisor

Prof. Mariel Vazquez

Prof. Wolfgang Polonik

Prof. Alex Aue

Rob Scharein

Lab's members

Stat. Department, UC Davis