Seriation, Spectral Clustering and de novo genome assembly

Antoine Recanati, *CNRS & ENS*

with Alexandre d'Aspremont, Thomas Kerdreux, Thomas Brüls,
CNRS - ENS Paris & Genoscope.

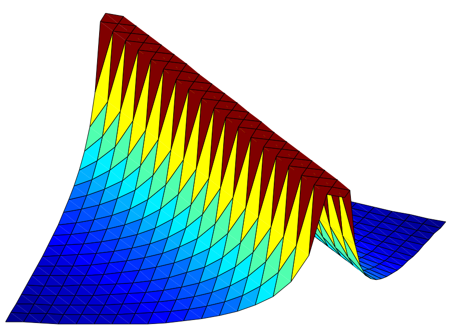
Seriation

The Seriation Problem.

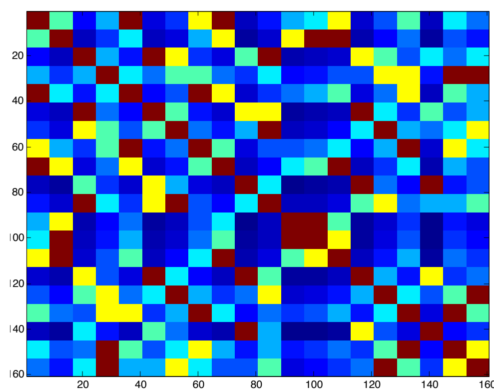
- Pairwise **similarity information** A_{ij} on n variables.
- Suppose the data has a **serial structure**, i.e. there is an order π such that

$$A_{\pi(i)\pi(j)} \text{ decreases with } |i-j| \quad \textbf{(R-matrix)}$$

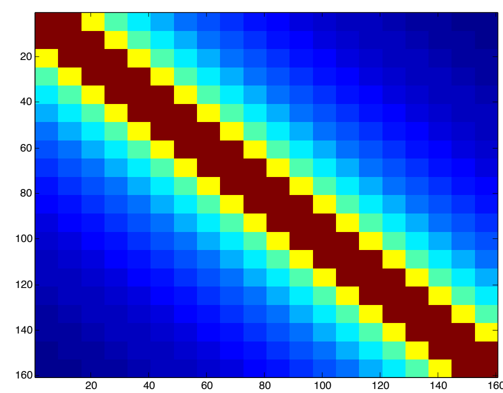
Recover π?



Similarity matrix



Input

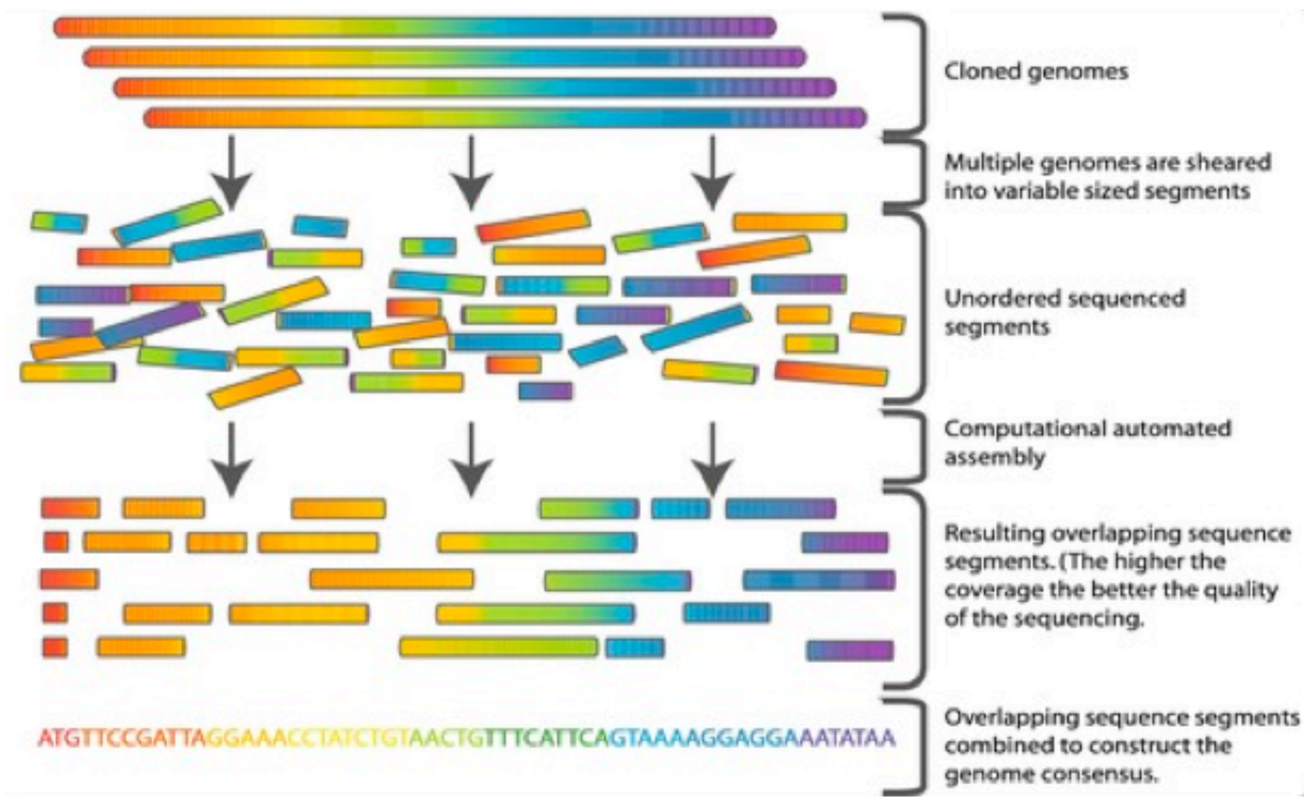


Reconstructed

Genome Assembly

Seriation has direct applications in *(de novo)* genome assembly.

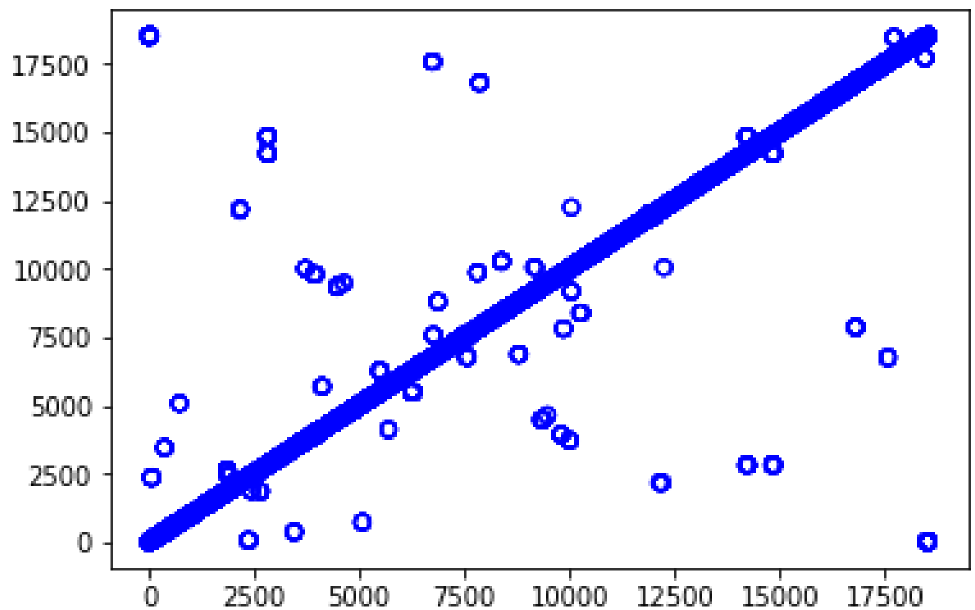
- Genomes are cloned multiple times and randomly cut into shorter reads ($\sim$ 400bp to 100kbp), which are fully sequenced.
- Reorder the reads to recover the genome.



Genome Assembly

Overlap Layout Consensus (OLC). Three stages.

- Compute **overlap** between all read pairs.
- **Reorder** overlap matrix to recover read order.
- Average the read values to create a **consensus** sequence.



The read reordering problem is a **seriation** problem.

Genome Assembly in Practice

Noise. In the noiseless case, the overlap matrix is a **R-matrix**. In practice. . .

- There are base calling **errors** in the reads, typically 2% to 20% depending on the process.
- Entire parts of the genome are **repeated**, which breaks the serial structure.

Sequencing technologies

- Next generation : short reads ($\sim$ 400bp), **few errors** ($\sim$ 2%). Repeats are challenging
- Third generation : **long reads** ($\sim$ 10kbp), more errors ($\sim$ 15%). Can resolve some repeats, but not all of them + noise can be challenging

Genome Assembly in Practice

Current assemblers.

- With **short accurate reads**, the reordering problem is solved by **combinatorial methods** using the topology of the assembly graph and additional pairing information.
- With **long noisy reads**, reads are **corrected** before assembly (hybrid correction or self-mapping).
- Layout and consensus not clearly separated, many **heuristics** . . .
- minimap+miniasm : first long raw reads straight assembler (but consensus sequence is as noisy as raw reads).

Outline

- Introduction
- **Spectral relaxation of Seriation (Spectral Ordering)**
- Multi-dimensional Spectral Ordering
- Results (Application to genome assembly)

2-SUM and the Graph Laplacian

The 2-SUM Combinatorial Problem.

- The **2-SUM problem** is written

$$\min_{\pi \in \mathcal{P}} \sum_{i,j=1}^{n} A_{\pi(i)\pi(j)} (i-j)^2$$

  or alternatively,

$$\min_{\pi \in \mathcal{P}} \sum_{i,j=1}^{n} A_{ij} (\pi(i) - \pi(j))^2$$

- optimal permutation π^* : high values of $A \Leftrightarrow$ low $|\pi(i) - \pi(j)|$, $i.e.$, i and j lay close to each other.

2-SUM and the Graph Laplacian

Graph Laplacian

- A : adjacency matrix of a undirected weighted graph ($A_{ij} > 0$ iff. there is an edge between nodes i and j).
- Node i has degree $d_i = \sum_j A_{ij}$. Degree matrix $D = \mathbf{diag}(A\mathbf{1}) = \mathbf{diag}(d)$.
- Laplacian matrix $L = D - A$.
- The **Laplacian** can be viewed as a **quadratic form**,

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} A_{ij} (f_i - f_j)^2$$

2-SUM and the Graph Laplacian

Mathematical reminder

- For a vector $f = (f_1, \dots, f_n)^T \in \mathbb{R}^n$ and a matrix $M \in \mathbb{R}^{n \times n}$, we have, $f^T M f = \sum_{i,j=1}^{n} M_{ij} f_i f_j$

- ($\lambda \in \mathbb{R}$, $u \in \mathbb{R}^n$) is a eigenvalue-eigenvector couple of $L \in \mathbb{R}^{n \times n}$ iff $Lu = \lambda u$

2-SUM and the Graph Laplacian

The **Laplacian** can be viewed as a **quadratic form**,

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} A_{ij} (f_i - f_j)^2$$

Indeed for any $f \in \mathbb{R}^n$,

$$\begin{aligned}
f^T L f = & \quad f^T D f - f^T A f \\
= & \quad \textstyle\sum_{i=1}^{n} f_i^2 D_{ii} - \sum_{i,j=1}^{n} A_{ij} f_i f_j \\
= & \quad \textstyle\sum_{i=1}^{n} f_i^2 (\sum_{j=1}^{n} A_{ij}) - \sum_{i,j=1}^{n} A_{ij} f_i f_j \\
= & \quad \textstyle\sum_{i,j=1}^{n} A_{ij} (f_i^2 - f_i f_j) \\
= & \quad \textstyle\frac{1}{2} \sum_{i,j=1}^{n} A_{ij} (f_j^2 + f_i^2 - 2 f_i f_j) \\
= & \quad \textstyle\frac{1}{2} \sum_{i,j=1}^{n} A_{ij} (f_i - f_j)^2
\end{aligned}$$

2-SUM and the Graph Laplacian

The **Laplacian** can be viewed as a **quadratic form**,

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} A_{ij} (f_i - f_j)^2$$

- L is symmetric and positive semi-definite.
- L has n non-negative, real-valued eigenvalues, $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.
- $\mathbf{1} = (1, \ldots, 1)^T$ is eigenvector associated to eigenvalue 0.
- If A has K connected components, the eigenvalue 0 has multiplicity $K + 1$, with eigenvectors being indicators of the connected components.
- If $f \in \{-1, +1\}^n$, objective of min-cut (clustering).

2-SUM and the Graph Laplacian

- The **2-SUM problem** is written

$$\min_{\pi \in \mathcal{P}} \quad \sum_{i,j=1}^{n} A_{\pi(i)\pi(j)} (i-j)^2$$

or alternatively,

$$\min_{\pi \in \mathcal{P}} \quad \sum_{i,j=1}^{n} A_{ij} (\pi(i) - \pi(j))^2$$

$i.e.$,

$$\min_{\pi \in \mathcal{P}} \quad \pi^T L \pi$$

- For certain matrices A, **2-SUM $\Longleftrightarrow$ seriation.** ([Fogel et al., 2013])

- **NP-Complete** for generic matrices A.

- Constraints $\pi \in \mathcal{P}$?

Spectral relaxation

$$\min_{\pi \in \mathcal{P}} \pi^T L_A \pi \qquad \text{(2SUM)}$$

Set of permutation vectors :

$$\begin{aligned} \pi(i) &\in \{1, ..., n\}, \quad \forall 1 \leq i \leq n \\ \pi^T \mathbf{1} &= n(n+1)/2 \\ \|\pi\|_2^2 &= n(n+1)(2n+1)/6 \end{aligned}$$

- Since $L\mathbf{1} = 0$, (2SUM) is invariant by $\pi \leftarrow \pi - \frac{(n+1)}{2}\mathbf{1}$, so enforce $\pi^T \mathbf{1} = 0$.
- Up to a dilatation, we can chose $\|\pi\|_2^2 = 1$.
- Relax the integer constraints and let $\pi(i) \in \mathbb{R}$.

Spectral relaxation

Spectral Seriation. Define the Laplacian of A as $L = \mathbf{diag}(A\mathbf{1}) - A$. The Fiedler vector of A is written

$$f = \underset{\substack{\mathbf{1}^T x = 0, \\ \|x\|_2 = 1}}{\operatorname{argmin}} \; x^T L_A x.$$

and is the second smallest eigenvector of the Laplacian.

The Fiedler vector reorders a R-matrix in the noiseless case.

Theorem [Atkins, Boman, and Hendrickson, 1998]

Spectral seriation. *Suppose $A \in \mathbf{S}_n$ is a pre-R matrix, with a simple Fiedler value whose Fiedler vector f has no repeated values. Suppose that $\Pi \in \mathcal{P}$ is such that the permuted Fiedler vector Πv is monotonic, then $\Pi A \Pi^T$ is an R-matrix.*

Spectral Ordering Algorithm

The Algorithm.

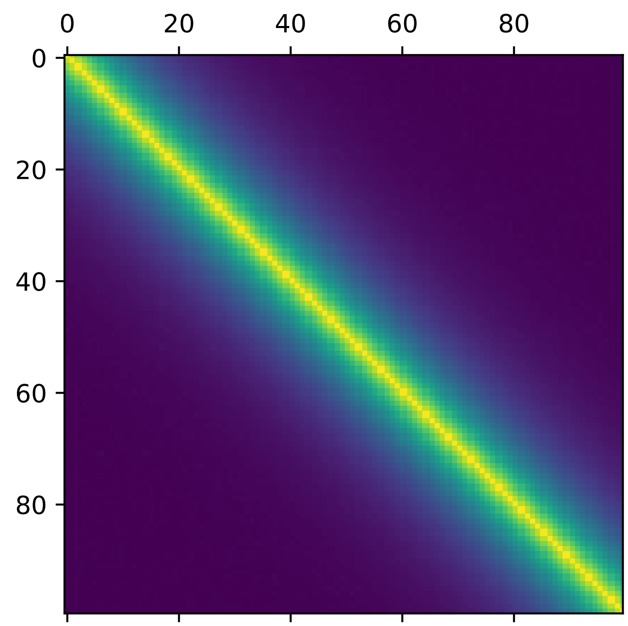
Input: Connected similarity matrix $A \in \mathbb{R}^{n \times n}$

1: Compute Laplacian $L = \mathbf{diag}(A\mathbf{1}) - A$

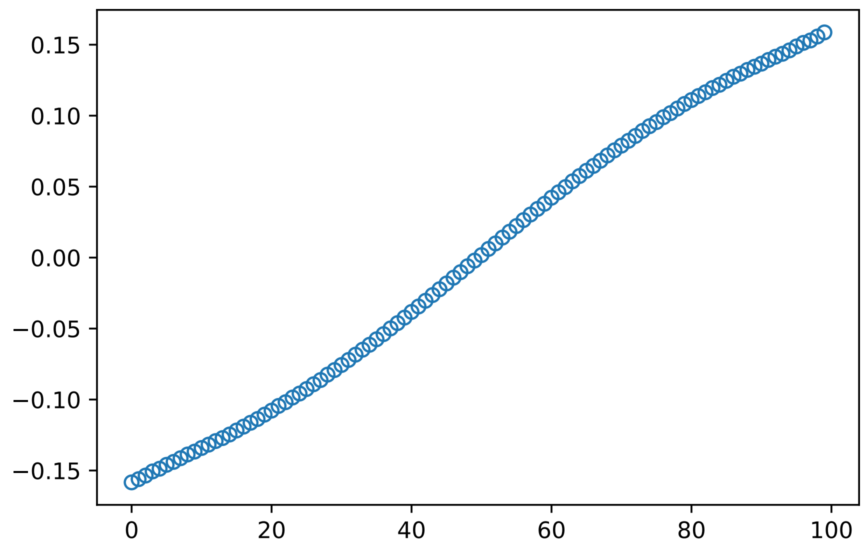
2: Compute second smallest eigenvector of L, $\mathbf{x}^*$

3: Sort the values of $\mathbf{x}^*$

Output: Permutation $\pi : \mathbf{x}^*{}_{\pi(1)} \leq \mathbf{x}^*{}_{\pi(2)} \leq \ldots \leq \mathbf{x}^*{}_{\pi(n)}$



Similarity matrix



Fiedler vector

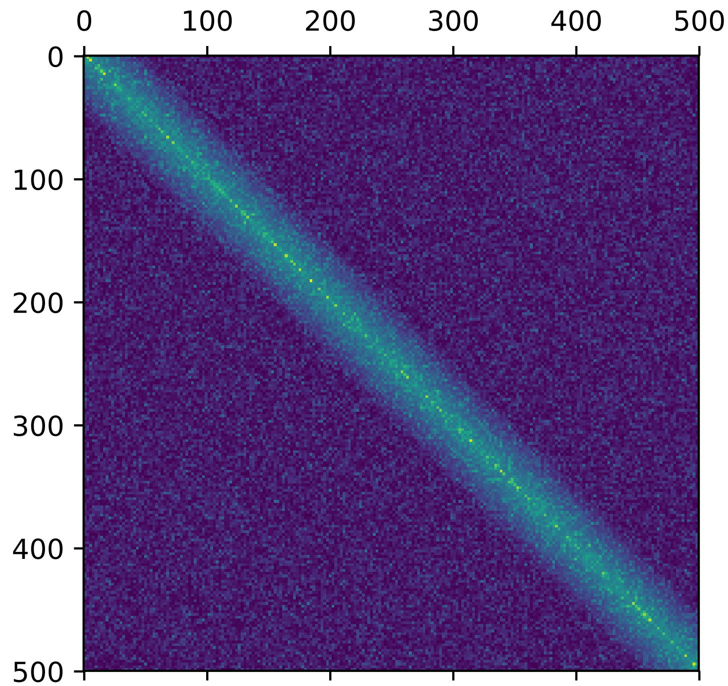
Spectral Solution

- Spectral solution easy to compute and scales well (polynomial time)
- But sensitive and not flexible (hard to include additional structural constraints)
- Other (convex) relaxations can handle structural constraints and solve more robust objectives than 2SUM

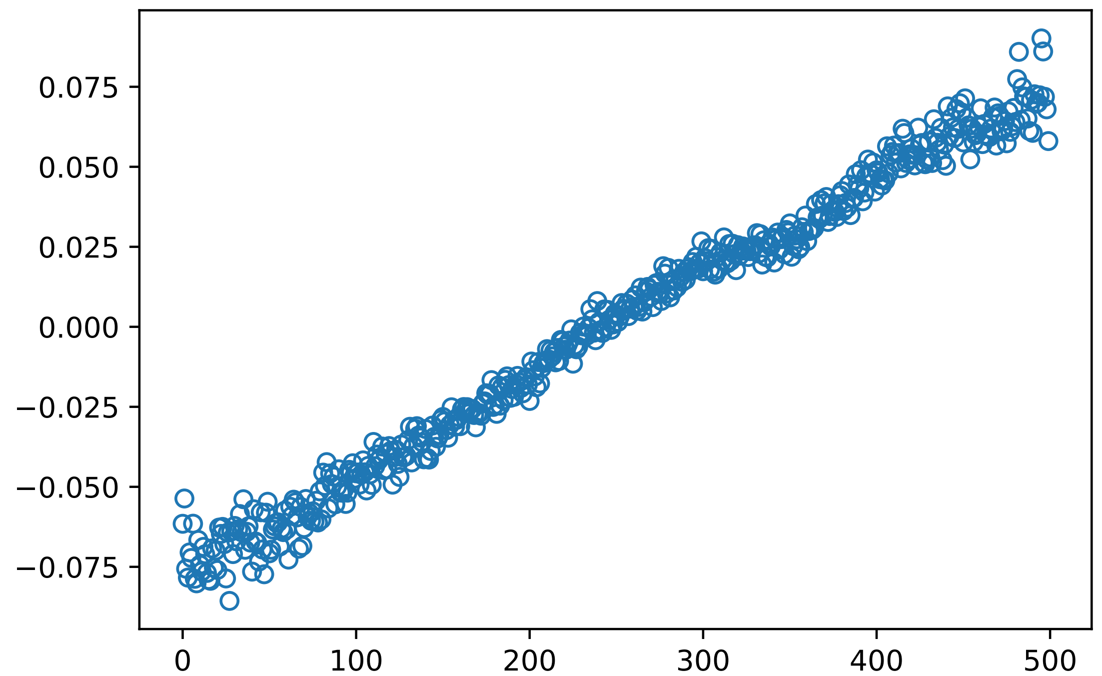
Genome assembly pipeline

- **Overlap** : computed from **k-mers**, yielding a similarity matrix A
- **Layout** : A is **thresholded** to remove noise-induced overlaps, and reordered with **spectral ordering algorithm**. Layout fine-grained with overlap information.
- **Consensus** : Genome sliced in windows

Spectral Solution vs Noisy Synthetic data



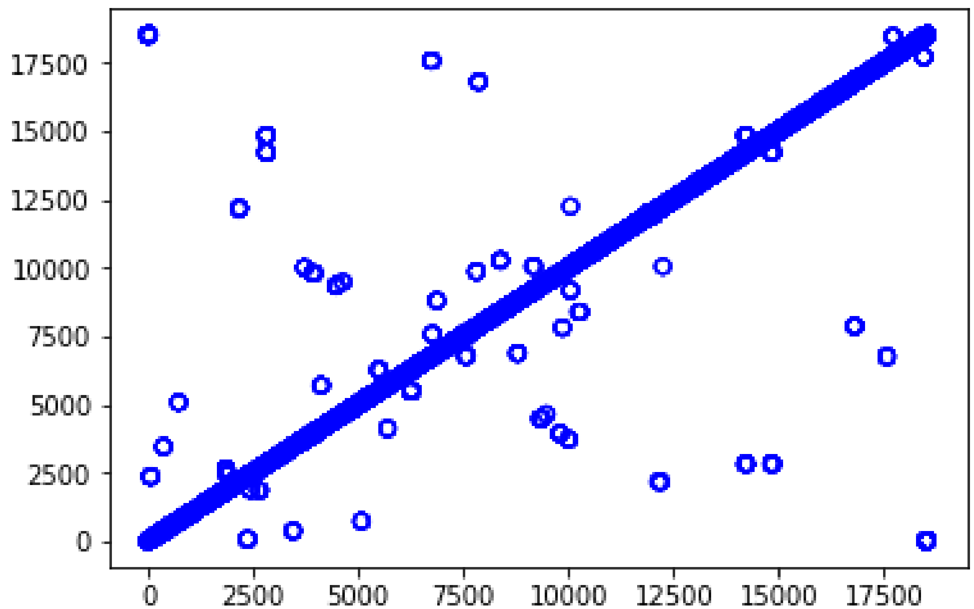
Similarity matrix



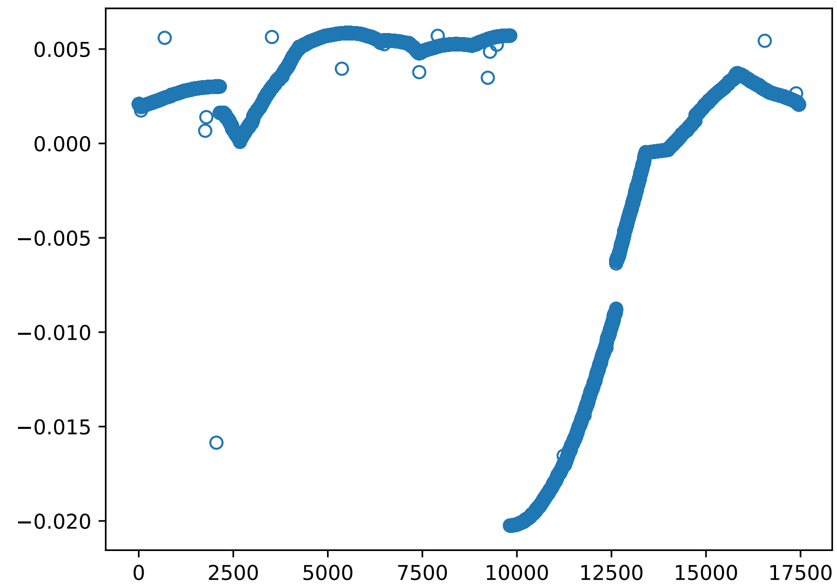
Fiedler vector

- Gaussian noise over perfect R-matrix.

Spectral Solution vs Real DNA data



Similarity matrix



Fiedler vector

- Repeats are a more structured noise that makes the method fail.

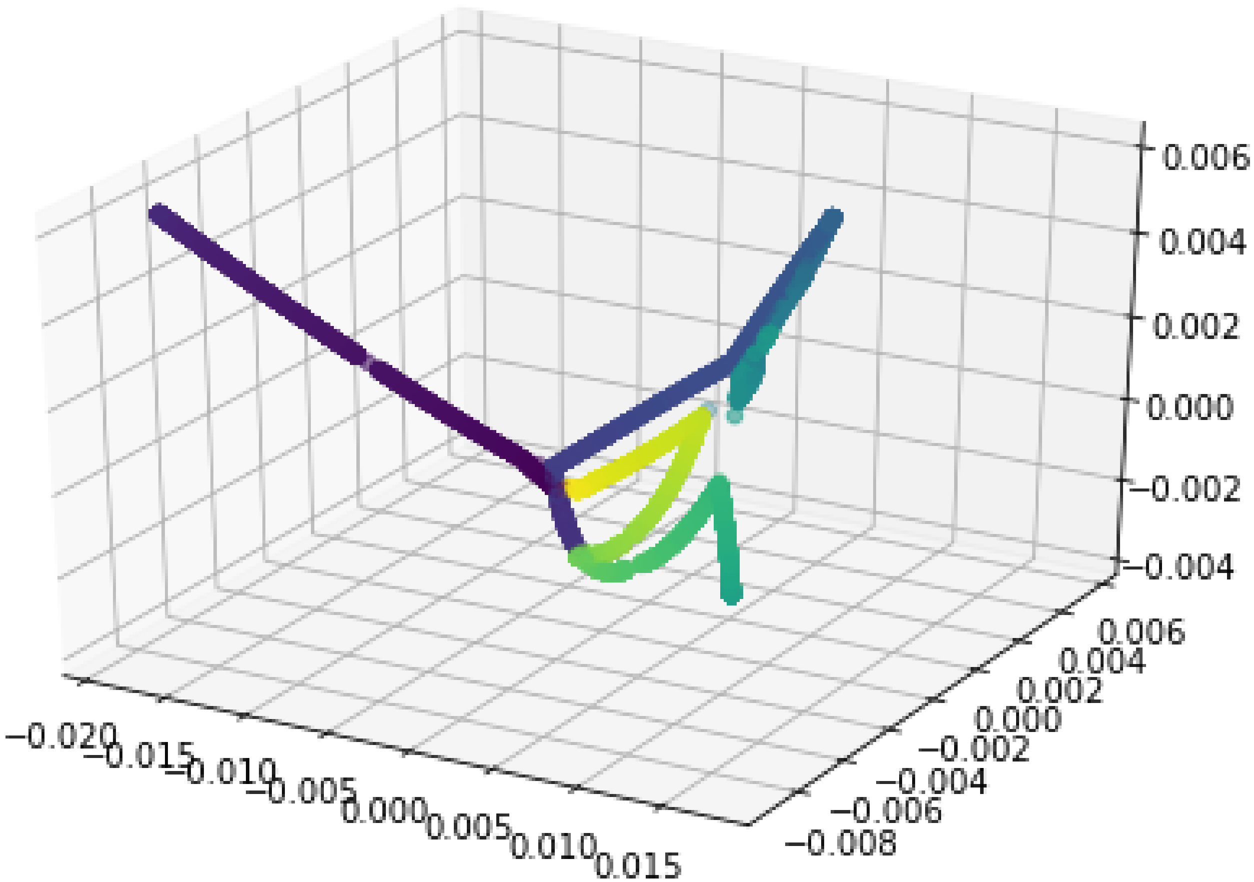
Outline

- Introduction
- Spectral relaxation of Seriation (Spectral Ordering)
- **Multi-dimensional Spectral Ordering**
- Results (Application to genome assembly)

Multi-dimensional Spectral Embedding

(Spoiler Alert!)

There is information in the rest of the eigenvectors of L



3d scatter plot of the 3 first non-zero eigenvectors of L

Multi-Dim 2-SUM and the Graph Laplacian

Generalize the quadratic expression involving the **Laplacian**,

$$\mathbf{Tr}\left(\tilde{\Phi}^T L_A \tilde{\Phi}\right) = \frac{1}{2} \sum_{i,j=1}^{n} A_{ij} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2$$

- Let $0 = \lambda_0 < \lambda_1 \leq \ldots \leq \lambda_{n-1}$, $\Lambda \triangleq \mathbf{diag}\left(\lambda_0, \ldots, \lambda_{n-1}\right)$, $\Phi = \left(\mathbf{1}, f_{(1)}, \ldots, f_{(n-1)}\right)$, be the eigendecomposition of $L = \Phi \Lambda \Phi^T$.
- For any $K < n$, $\Phi^{(K)} \triangleq \left(f_{(1)}, \ldots, f_{(K)}\right)$ defines a K-dimensional embedding

$$\boldsymbol{y}_i = \left(f_{(1)}(i), f_{(2)}(i), \ldots, f_{(K)}(i)\right)^T \in \mathbb{R}^K, \quad \text{for} \quad i = 1, \ldots, n. \tag{K-LE}$$

which solves the following embedding problem,

$$\begin{array}{ll} \text{minimize} & \sum_{i,j=1}^{n} A_{ij} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2 \\ \text{such that} & \tilde{\Phi} = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T\right)^T \in \mathbb{R}^{n \times K}, \ \tilde{\Phi}^T \tilde{\Phi} = \mathbf{I}_K, \ \tilde{\Phi}^T \mathbf{1}_n = \mathbf{0}_K \end{array} \tag{Lap-Emb}$$

Intermission : Spectral Clustering

Spectral Clustering usually leverages the first few eigenvectors of L. To partition data in K clusters,

- Compute the K lowest non-zero eigenvectors of L, $\Phi^{(K)} = \left(f_{(1)}, \ldots, f_{(K)}\right) \in \mathbb{R}^{n \times K}$.
- Run the K-means algorithm on this K-dimensional embedding.

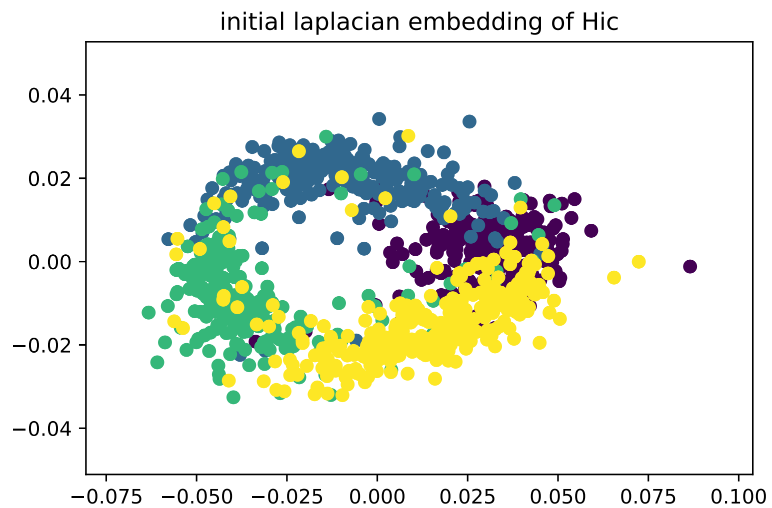
Multi-Dimensional Spectral Ordering (MDSO)

How to extract ordering from multidimensional embedding ?

- Construct new similarity matrix S
- For each point u, take k-NN in the embedding, fit by a line, use projection on the line to define similarity S_{ij} between points $i, j \in \text{kNN}(u)$.
- Run basic Spectral Ordering on S.
- If S is not connected, reorder each connected component, and use A to merge the ends.

Multi-Dimensional Spectral Ordering (MDSO)

- Simple generalization of Spectral Ordering
- Acts like a pre-preocessing on the similarity matrix
- Improves robustness to noise
- Handles circular orderings (with 2D embedding in a circle)



2D spectral embedding from similarity between single-cell Hi-C contact matrices

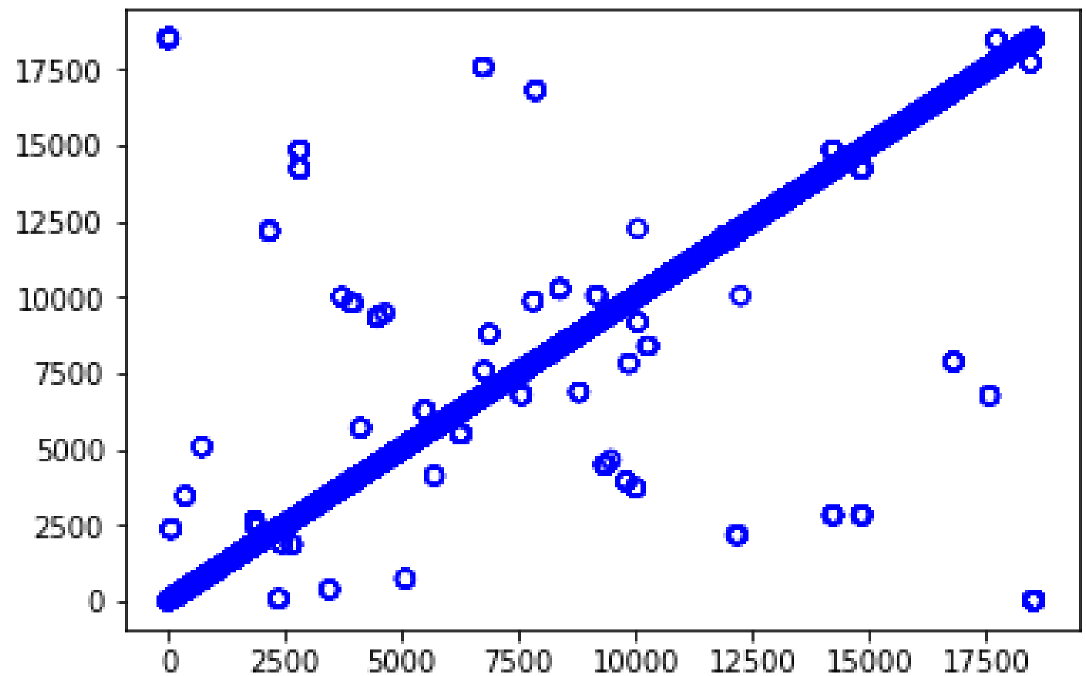
Outline

- Introduction
- Spectral relaxation of Seriation (Spectral Ordering)
- Multi-dimensional Spectral Ordering
- **Results (Application to genome assembly)**

Application to genome assembly

Bacterial genome.

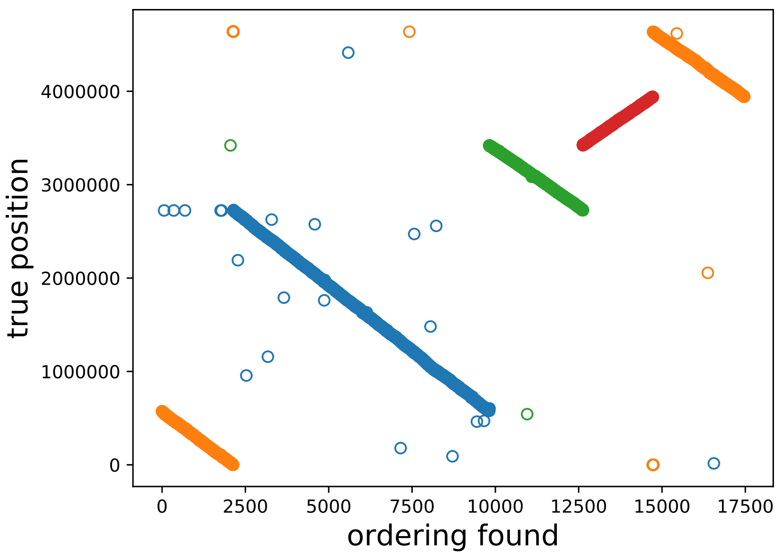
- *Escherichia coli* reads sequenced by Loman et al. [2015]. $\sim$ 4Mbp
- Oxford Nanopore Technology MinION's device (third generation long reads).
- minimap2 used to compute overlap-based similarity between reads.



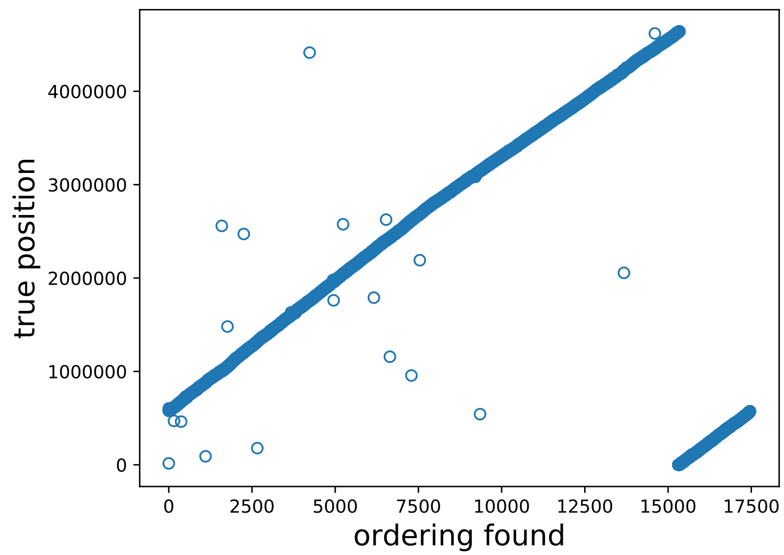
Application to genome assembly

Layout.

- MDSO new similarity matrix S is disconnected.
- Connected components can be merged by looking at the similarity between their ends from the original matrix A.
- Kendall-Tau score with reference ordering : 99.5%
- Full assembly pipeline yields $\sim$ 99% avg. identity (using MSA in sliding window)



Order in connected components



Merged ordering

Conclusion

- Equivalence **2-SUM** $\Longleftrightarrow$ **seriation.**
- **Spectral ordering** : simple relaxation of 2-SUM using spectrum of the Laplacian. Related to widespread Spectral Clustering algorithm.
- **Spectral ordering** is sensitive to **repeats**.
- **Multi-dimensional Spectral Ordering** can overcome this issue (and solve circular seriation).
- **Straightforward assembly pipeline** with MSA to perform consensus.

*

References

Jonathan E Atkins, Erik G Boman, and Bruce Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.

Fajwel Fogel, Rodolphe Jenatton, Francis Bach, and Alexandre d'Aspremont. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems*, pages 1016–1024, 2013.

Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, 12(8):733, 2015.

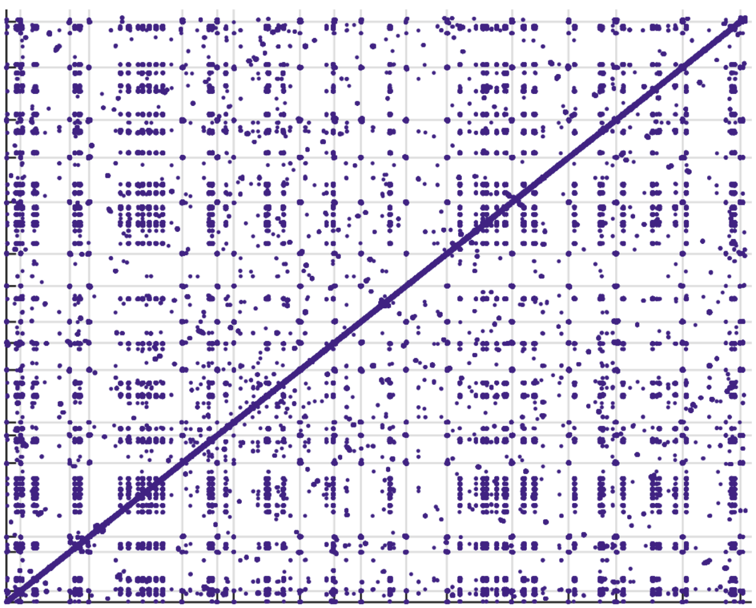
Antoine Recanati, Thomas Brüls, and Alexandre dAspremont. A spectral algorithm for fast de novo layout of uncorrected long nanopore reads. *Bioinformatics*, 33(20):3188–3194, 2017.

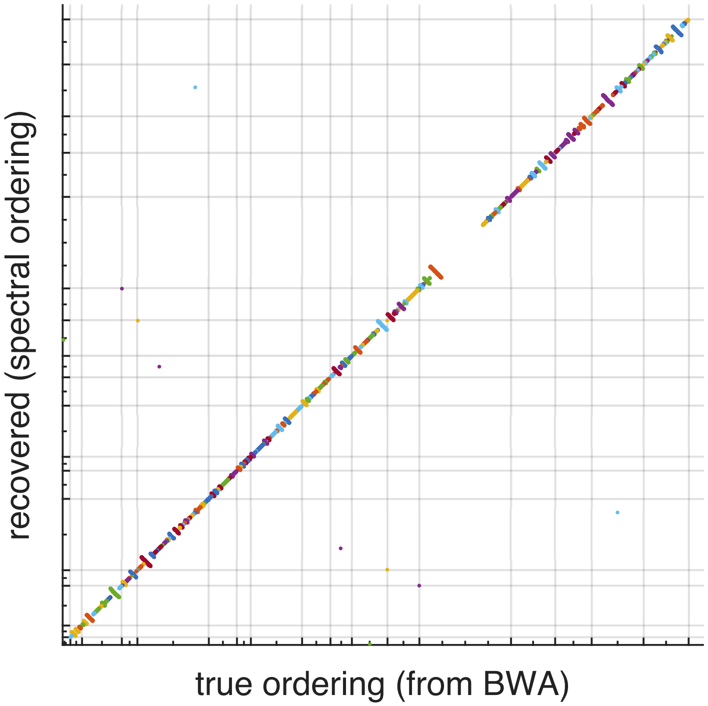
Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Application to genome assembly

Eukaryotic genome : $\boldsymbol{S.\ Cerevisiae}$

- 16 chromosomes
- Many repeats
- Higher threshold on similarity matrix $\Rightarrow$ many connected components





Conclusion

Straightforward assembly pipeline.

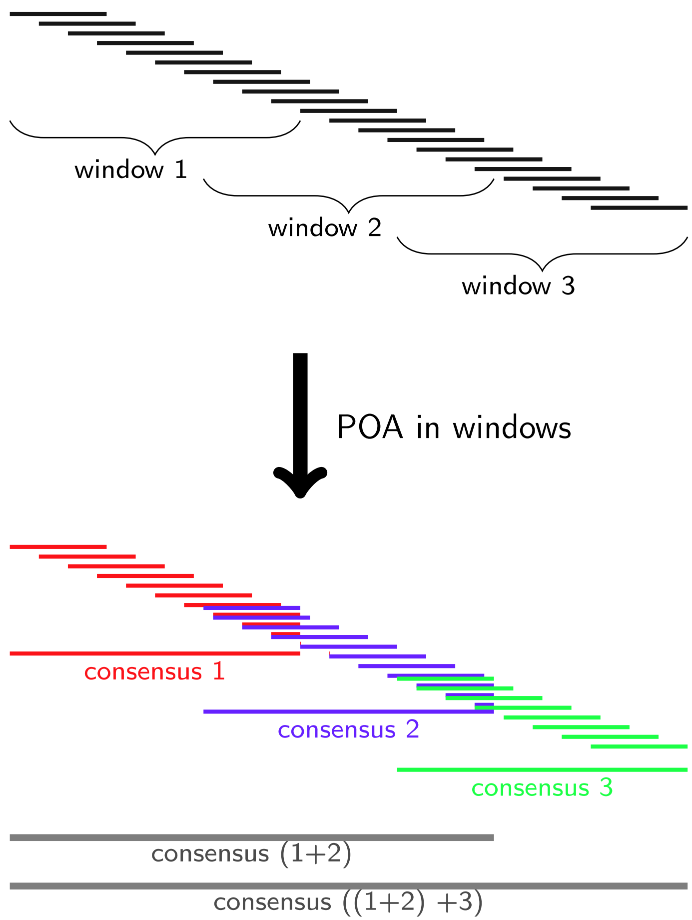
- Equivalence **2-SUM** $\Longleftrightarrow$ **seriation.**
- **Layout** correctly found by **spectral relaxation** for **bacterial genomes** (with limited number of repeats)
- **Consensus** computed by **MSA** in sliding windows $\Rightarrow\sim$ 99% avg. identity with reference

Future work.

- **Additional information** could help assemble more **complex genomes** (e.g. with topological constraints on the similarity graph, or chromosome assignment...)
- Other problems involving Seriation ?
- **Convex relaxations** can also handle **constraints** (e.g. $|\pi(i) - \pi(j)| \leq k$) for different problems

Consensus

- Once layout is computed and fined-grained, slicing in windows
- Multiple Sequence Alignment using Partial Order Graphs (POA) in windows
- Windows merging



Combinatorial problems.

- The **2-SUM problem** is written

$$\min_{\pi \in \mathcal{P}} \sum_{i,j=1}^{n} A_{\pi(i)\pi(j)}(i-j)^2 \quad \text{or equivalently} \quad \min_{\pi \in \mathcal{P}} \pi^T L_A \pi$$

where L_A is the Laplacian of A.

- **NP-Complete** for generic matrices A.

Convex Relaxation

Seriation as an optimization problem.

$$\min_{\pi \in \mathcal{P}} \sum_{i,j=1}^{n} A_{\pi(i)\pi(j)}(i-j)^2$$

What's the point?

- Gives a spectral (hence polynomial) solution for 2-SUM on some R-matrices.
- Write a **convex relaxation** for 2-SUM and seriation.
  - Spectral solution scales very well (cf. Pagerank, spectral clustering, etc.)
  - Not very robust. . .
  - Not flexible. . . Hard to include additional structural constraints.

Convex Relaxation

- Let $\mathcal{D}_n$ the set of doubly stochastic matrices, where

$$\mathcal{D}_n = \{X \in \mathbb{R}^{n\times n} : X \geqslant 0, X\mathbf{1} = \mathbf{1}, X^T\mathbf{1} = \mathbf{1}\}$$

is the **convex hull of the set of permutation matrices.**

- Notice that $\mathcal{P} = \mathcal{D} \cap \mathcal{O}$, i.e. Π permutation matrix if and only Π is both **doubly stochastic** and **orthogonal.**

- Solve

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(Y^T\Pi^T L_A \Pi Y) - \mu\|P\Pi\|_F^2 \\ \text{subject to} & e_1^T\Pi g + 1 \leq e_n^T \Pi g, \\ & \Pi\mathbf{1} = \mathbf{1},\ \Pi^T\mathbf{1} = \mathbf{1}, \\ & \Pi \geq 0, \end{array} \qquad (1)$$

in the variable $\Pi \in \mathbb{R}^{n\times n}$, where $P = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and $Y \in \mathbb{R}^{n\times p}$ is a matrix whose columns are small perturbations of $g = (1, \ldots, n)^T$.

Convex Relaxation

Objective. $\mathbf{Tr}(Y^T\Pi^T L_A \Pi Y) - \mu\|P\Pi\|_F^2$

- **2-SUM** term $\mathbf{Tr}(Y^T\Pi^T L_A \Pi Y) = \sum_{i=1}^p y_i^T \Pi^T L_A \Pi y_i$ where y_i are small perturbations of the vector $g = (1, \ldots, n)^T$.
- **Orthogonalization penalty** $-\mu\|P\Pi\|_F^2$, where $P = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.
    - Among all DS matrices, rotations (hence permutations) have the highest Frobenius norm.
    - Setting $\mu \leq \lambda_2(L_A)\lambda_1(YY^T)$, keeps the problem **a convex QP.**

Constraints.

- $e_1^T\Pi g + 1 \leq e_n^T \Pi g$ breaks degeneracies by imposing $\pi(1) \leq \pi(n)$. Without it, both monotonic solutions are optimal and this degeneracy can significantly deteriorate relaxation performance.
- $\Pi\mathbf{1} = \mathbf{1}$, $\Pi^T\mathbf{1} = \mathbf{1}$ and $\Pi \geq 0$, keep Π doubly stochastic.

Convex Relaxation

Other relaxations.

- Relaxations for orthogonality constraints, e.g. SDPs in [**???**]. Simple idea: $Q^TQ = \mathbf{I}$ is a quadratic constraint on Q, **lift it.** This yields a $O(\sqrt{n})$ approximation ratio.
- $O(\sqrt{\log n})$ approximation bounds for **Minimum Linear Arrangement** [**??????**].
- All these relaxations form extremely large SDPs.

Our simplest relaxation is a QP. No approximation bounds at this point however.

Convex Relaxation.

- **Semi-Supervised Seriation.** We can add structural constraints to the relaxation, where

$$a \leq \pi(i) - \pi(j) \leq b \quad \text{is written} \quad a \leq e_i^T \Pi g - e_j^T \Pi g \leq b.$$

  which are linear constraints in Π.

- **Sampling permutations.** We can generate permutations from a doubly stochastic matrix D
  - Sample monotonic random vectors u.
  - Recover a permutation by reordering Du.

- **Algorithms.** Large QP, projecting on doubly stochastic matrices can be done very efficiently, using block coordinate descent on the dual. Extended formulations by [**?**] can reduce the dimension of the problem to $O(n \log n)$ [**?**].

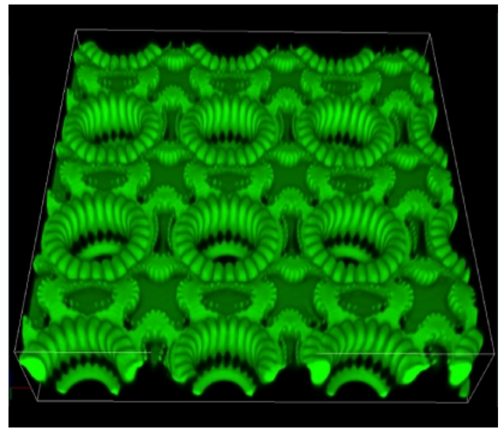
Numerical results: nanopores

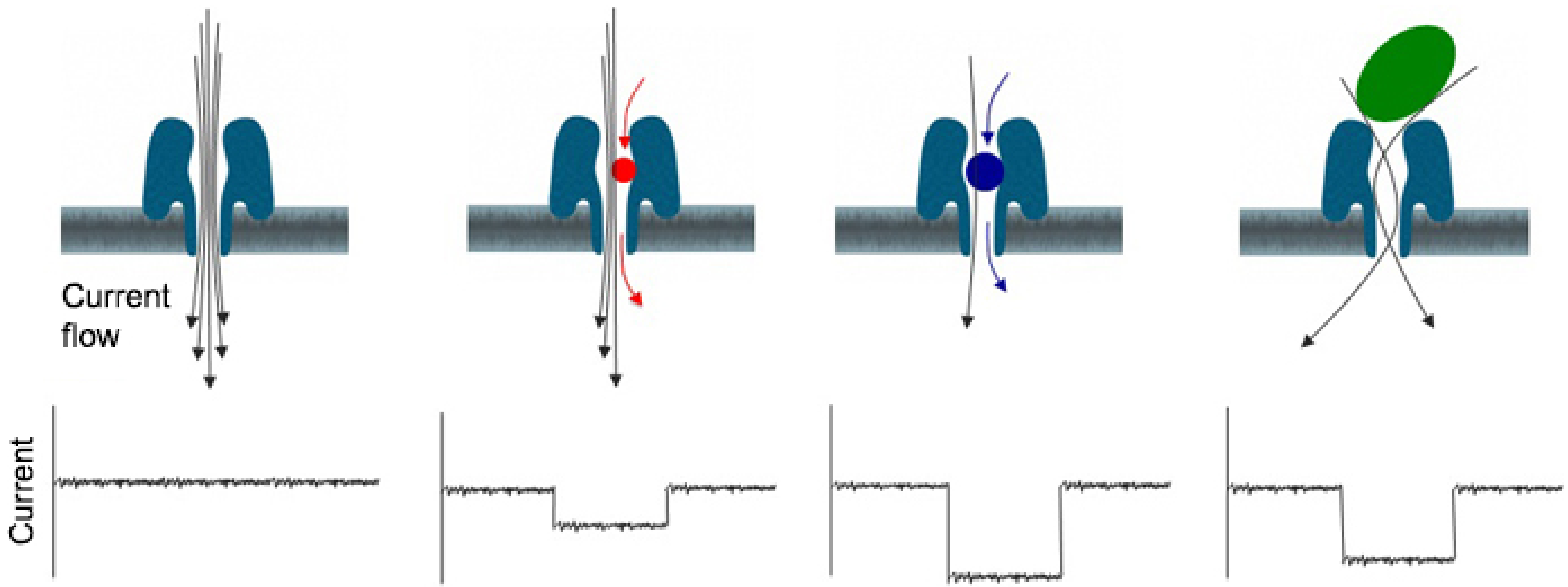
Nanopores DNA data. New sequencing hardware.

Oxford nanopores MinION.

Numerical results: nanopores

Nanopores.





Numerical results

Nanopores DNA data.

- **Longer reads.** Average 10k base pairs in early experiments. Compared with ~ 100 base pairs for existing technologies.
- **High error rate.** About 20% compared with a few percents for existing technologies.
- **Real-time data.** Sequencing data flows continuously.