Seriation and de novo genome assembly

Antoine Recanati, *CNRS & ENS*

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

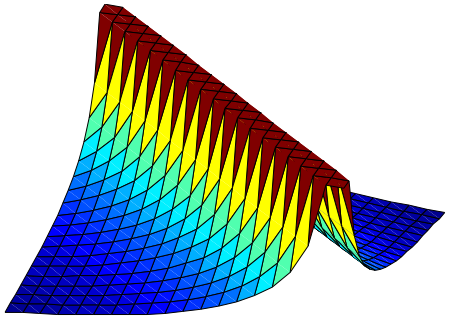
Seriation

The Seriation Problem.

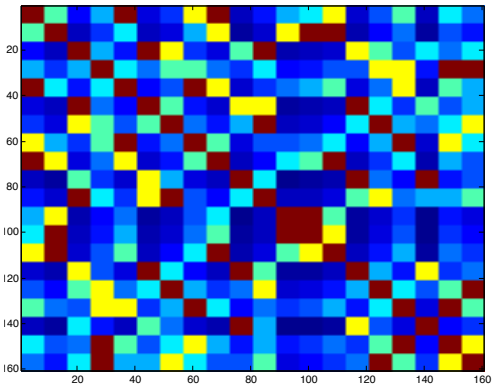
- Pairwise **similarity information** A_{ij} on n variables.
- Suppose the data has a **serial structure**, i.e. there is an order π such that

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

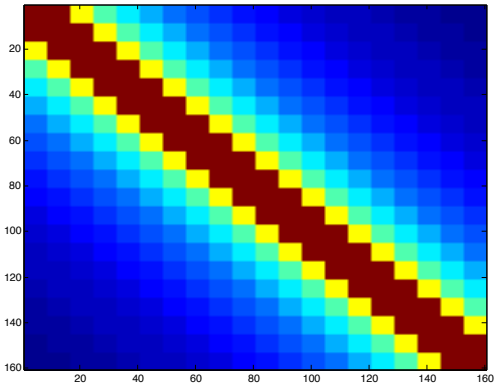
Recover π?



Similarity matrix



Input

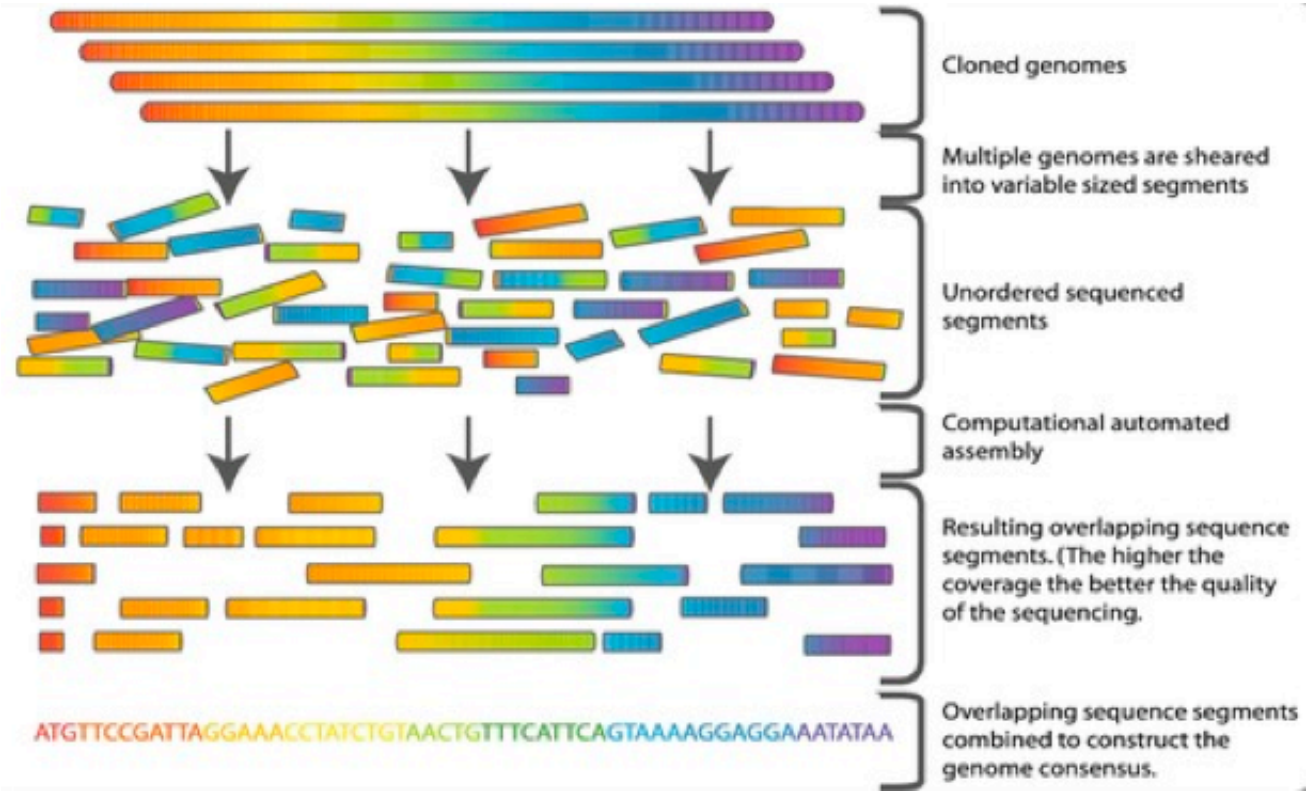


Reconstructed

Genome Assembly

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

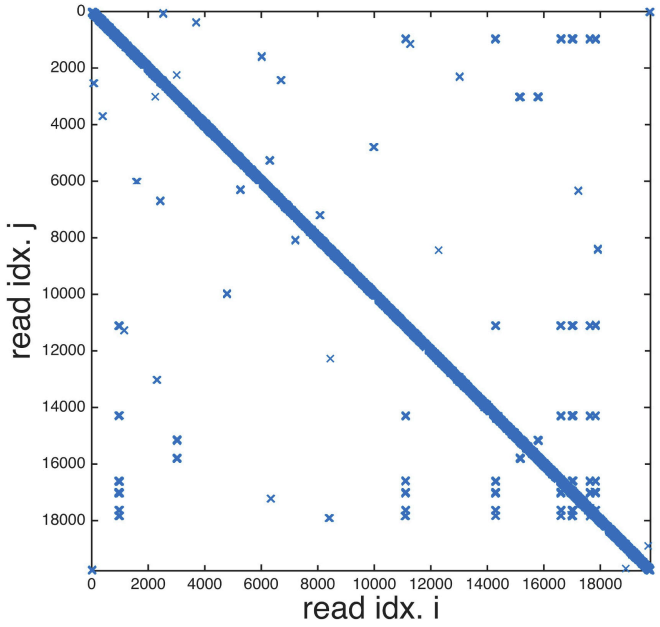
- Genomes are cloned multiple times and randomly cut into shorter reads ($\sim$ 400bp to 10kbp), which are fully sequenced.
- Reorder the reads to recover the genome.



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 repeats, but noise is 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** . . .
- miniasm : first long raw reads straight assembler (but consensus sequence is as noisy as raw reads).

Outline

- Introduction
- **Combinatorial problem**
- Spectral relaxation
- Results (Application to genome assembly)

Combinatorial problem (2-SUM)

2-SUM.

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

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

- Define $L_A = \mathbf{diag}(A\mathbf{1}) - A$ is the Laplacian of A. The 2-SUM problem is equivalently written

$$\min_{\pi \in \mathcal{P}} \ \pi^T L_A \pi$$

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

$$\begin{aligned}
x^T L_A x = & \quad x^T \mathbf{diag}(A\mathbf{1})x - x^T A x \\
= & \quad \textstyle\sum_{i=1}^n x_i^2 (\sum_{j=1}^n A_{ij}) - \sum_{i,j=1}^n A_{ij} x_i x_j \\
= & \quad \textstyle\sum_{i,j=1}^n A_{ij}(x_i^2 - x_i x_j) \\
= & \quad \textstyle\frac{1}{2}\sum_{i,j=1}^n A_{ij}(x_j^2 + x_i^2 - 2x_i x_j) \\
= & \quad \textstyle\frac{1}{2}\sum_{i,j=1}^n A_{ij}(x_i - x_j)^2
\end{aligned}$$

Seriation and 2-SUM

Combinatorial Solution.

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

Spectral relaxation

2-SUM problem :

$$\min_{\pi \in \mathcal{P}} \ \pi^T L_A \pi$$

NP-Complete for generic matrices A.

Set of permutation vectors :

$$\pi_i \in \{1, ..., n\}, \qquad \forall 1 \leq i \leq n$$
$$\pi^T \mathbf{1} = \tfrac{n(n+1)}{2}$$
$$\|\pi\|_2^2 = \tfrac{n(n+1)(2n+1)}{6}$$

Let $c = \frac{n+1}{2}\mathbf{1}$. $L_A \mathbf{1} = 0$. Withdrawing c from any vector π does not change the objective value. Up to a constant factor, the Fiedler vector f defined as follows solves a continuous relaxation of 2-SUM

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

Spectral relaxation

Spectral Seriation. Define the Laplacian of A as $L_A = \mathbf{diag}(A\mathbf{1}) - A$, 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, Hendrickson, et al., 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 Fielder vector Πv is monotonic, then $\Pi A \Pi^T$ is an R-matrix.*

Spectral Solution

- Spectral solution easy to compute and scales well
- But sensitive and not flexible (hard to include additional structural constraints)
- Other (convex) relaxations handle structural constraints

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

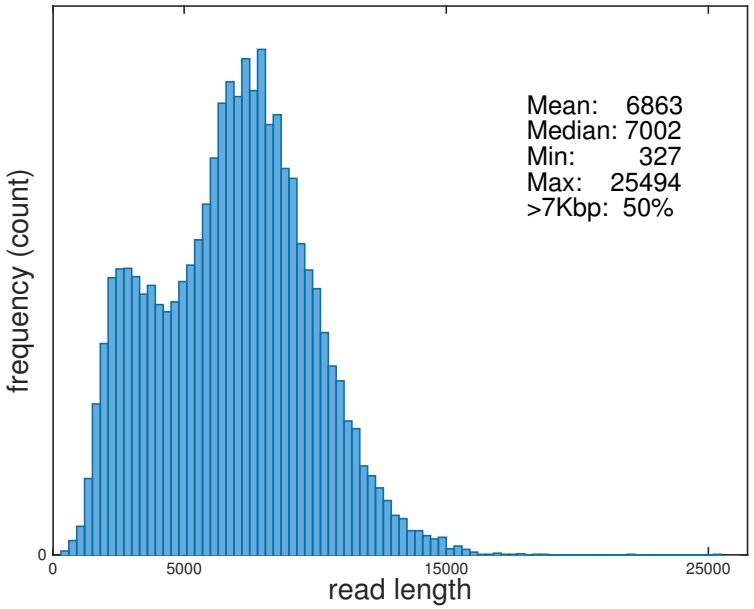
Outline

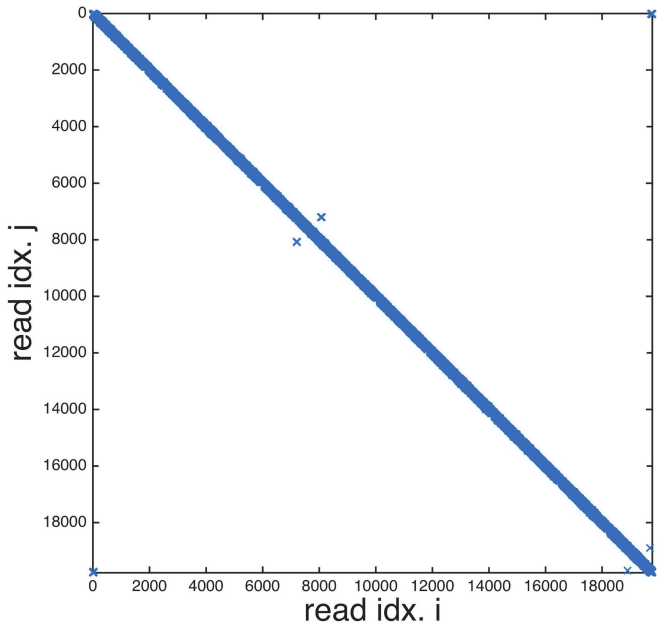
- Introduction
- Combinatorial problem
- Spectral relaxation
- **Results (Application to genome assembly)**

Application to genome assembly

Bacterial genomes.

- Long raw reads (Oxford Nanopore Technology)
- Overlaps computed with minimap : hashing k-mers
- Threshold on similarity matrix to remove false-overlaps

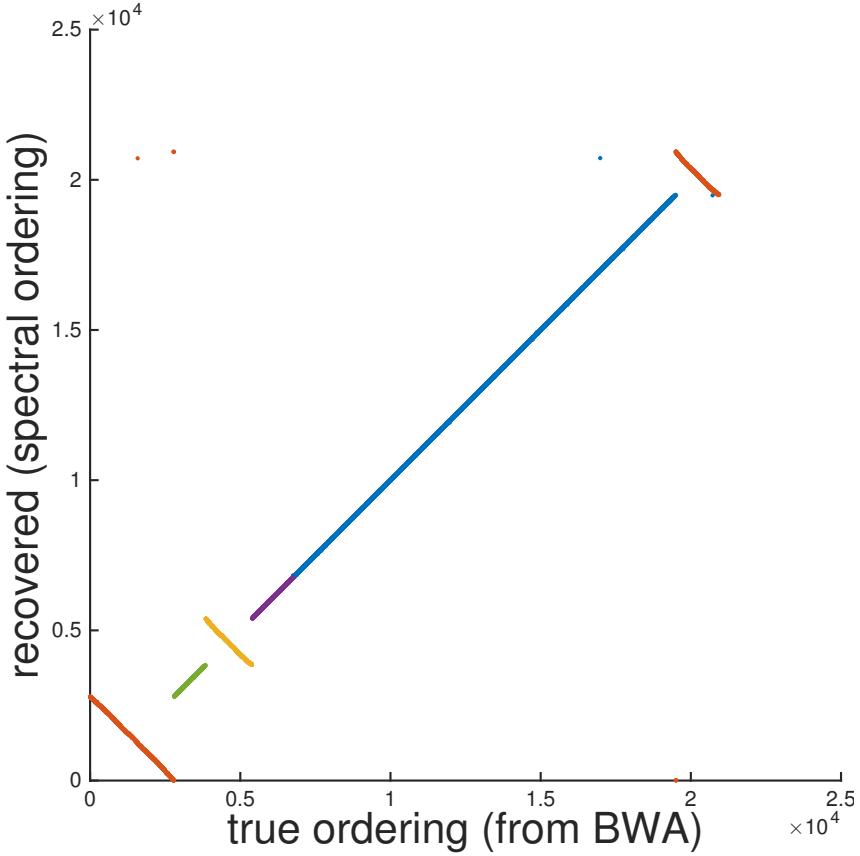




Application to genome assembly

Layout.

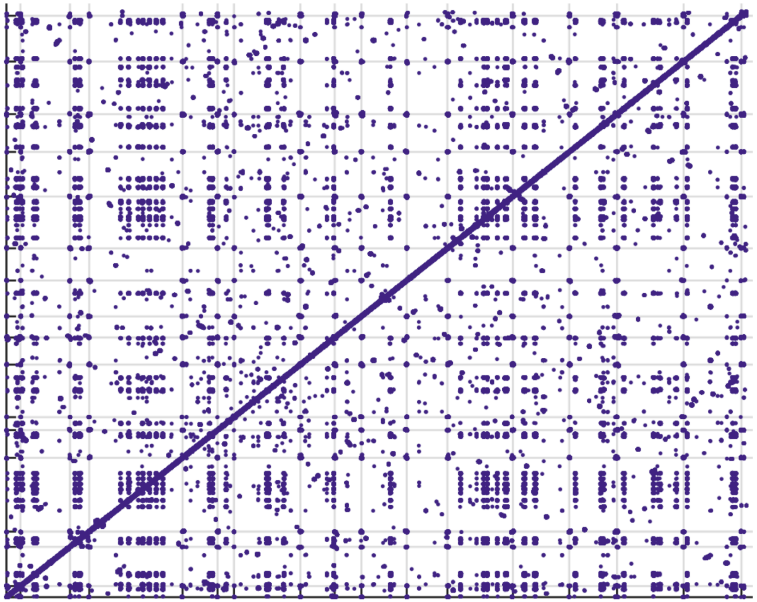
- Two bacterial genomes : $E.\ Coli$ and $A.\ Baylyi$
- Circular genomes, size $\sim$ 4Mbp
- A few connected components after threshold

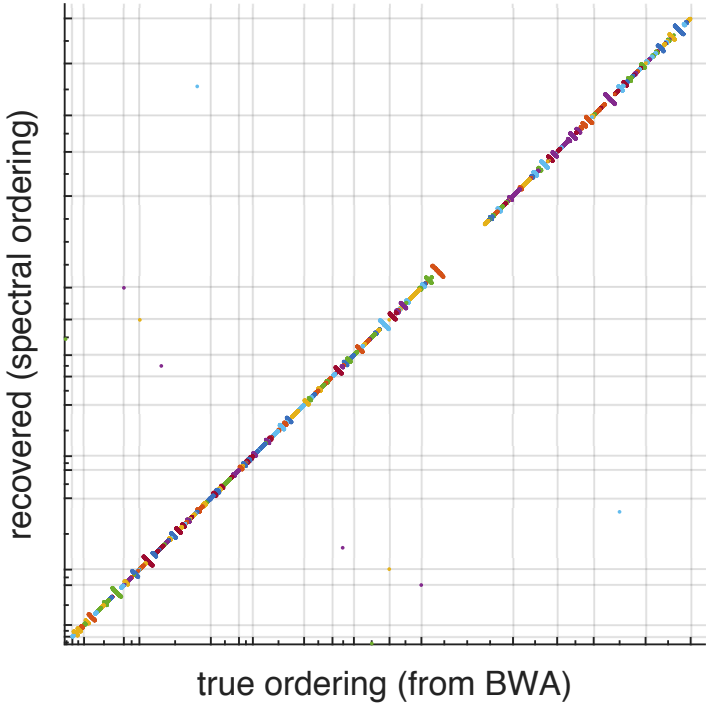


Application to genome assembly

Eukaryotic genome : *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

*

References

J.E. Atkins, E.G. Boman, B. Hendrickson, et al. A spectral algorithm for seriation and the consecutive ones problem. *SIAM J. Comput.*, 28 (1):297–310, 1998.

Avrim Blum, Goran Konjevod, R Ravi, and Santosh Vempala. Semidefinite relaxations for minimum bandwidth and other vertex ordering problems. *Theoretical Computer Science*, 235(1):25–42, 2000.

Moses Charikar, Mohammad Taghi Hajiaghayi, Howard Karloff, and Satish Rao. l_2^2 spreading metrics for vertex ordering problems. *Algorithmica*, 56(4):577–604, 2010.

R. Coifman, Y. Shkolnisky, F.J. Sigworth, and A. Singer. Cryo-EM structure determination through eigenvectors of sparse matrices. *working paper*, 2008.

Guy Even, Joseph Seffi Naor, Satish Rao, and Baruch Schieber. Divide-and-conquer approximation algorithms via spreading metrics. *Journal of the ACM (JACM)*, 47(4):585–616, 2000.

Uriel Feige. Approximating the bandwidth via volume respecting embeddings. *Journal of Computer and System Sciences*, 60(3):510–539, 2000.

Uriel Feige and James R Lee. An improved approximation ratio for the minimum linear arrangement problem. *Information Processing Letters*, 101(1):26–29, 2007.

F. Fogel, R. Jenatton, F. Bach, and A. d'Aspremont. Convex relaxations for permutation problems. *NIPS 2013, arXiv:1306.4805*, 2013.

Michel X. Goemans. Smallest compact formulation for the permutahedron. *Mathematical Programming*, pages 1–7, 2014.

David G Kendall. Incidence matrices, interval graphs and seriation in archeology. *Pacific Journal of mathematics*, 28(3):565–570, 1969.

Cong Han Lim and Stephen J Wright. Beyond the birkhoff polytope: Convex relaxations for vector permutation problems. *arXiv preprint arXiv:1407.6609*, 2014.

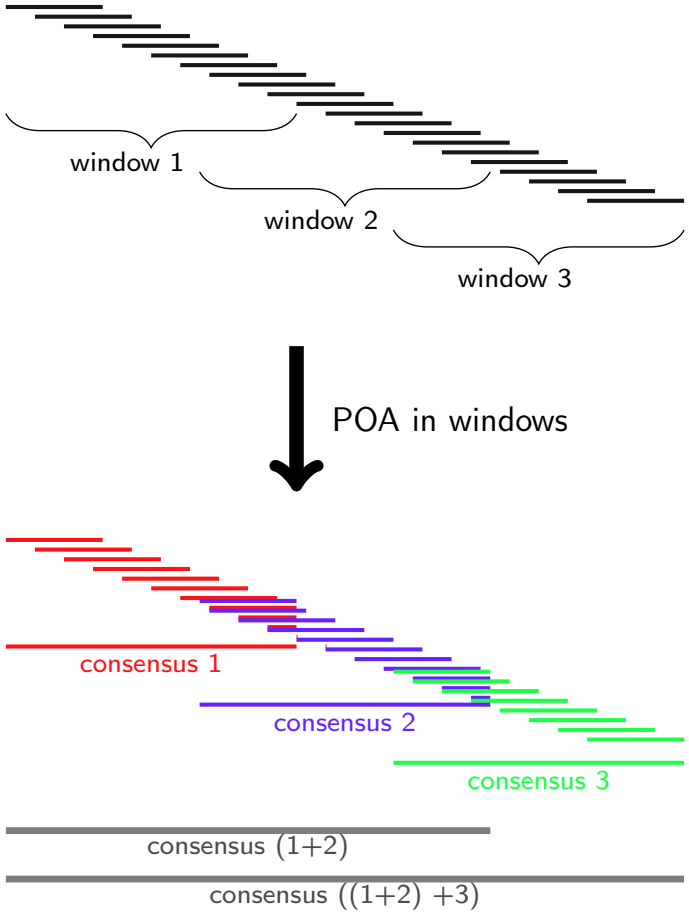
A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Mathematical programming*, 109(2):283–317, 2007.

Satish Rao and Andréa W Richa. New approximation techniques for some linear ordering problems. *SIAM Journal on Computing*, 34(2): 388–404, 2005.

Anthony Man-Cho So. Moment inequalities for sums of random matrices and their applications in optimization. *Mathematical programming*, 130(1):125–151, 2011.

Consensus

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



Seriation

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

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

  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 [Nemirovski, 2007, Coifman et al., 2008, So, 2011]. 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** [Even et al., 2000, Feige, 2000, Blum et al., 2000, Rao and Richa, 2005, Feige and Lee, 2007, Charikar et al., 2010].
- All these relaxations form extremely large SDPs.

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

Semi-Supervised Seriation

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 [Goemans, 2014] can reduce the dimension of the problem to $O(n \log n)$ [Lim and Wright, 2014].

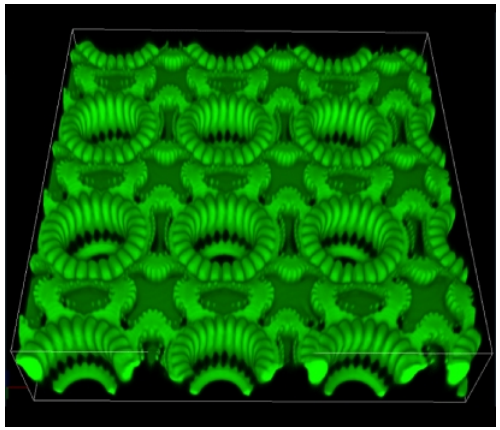
Numerical results: nanopores

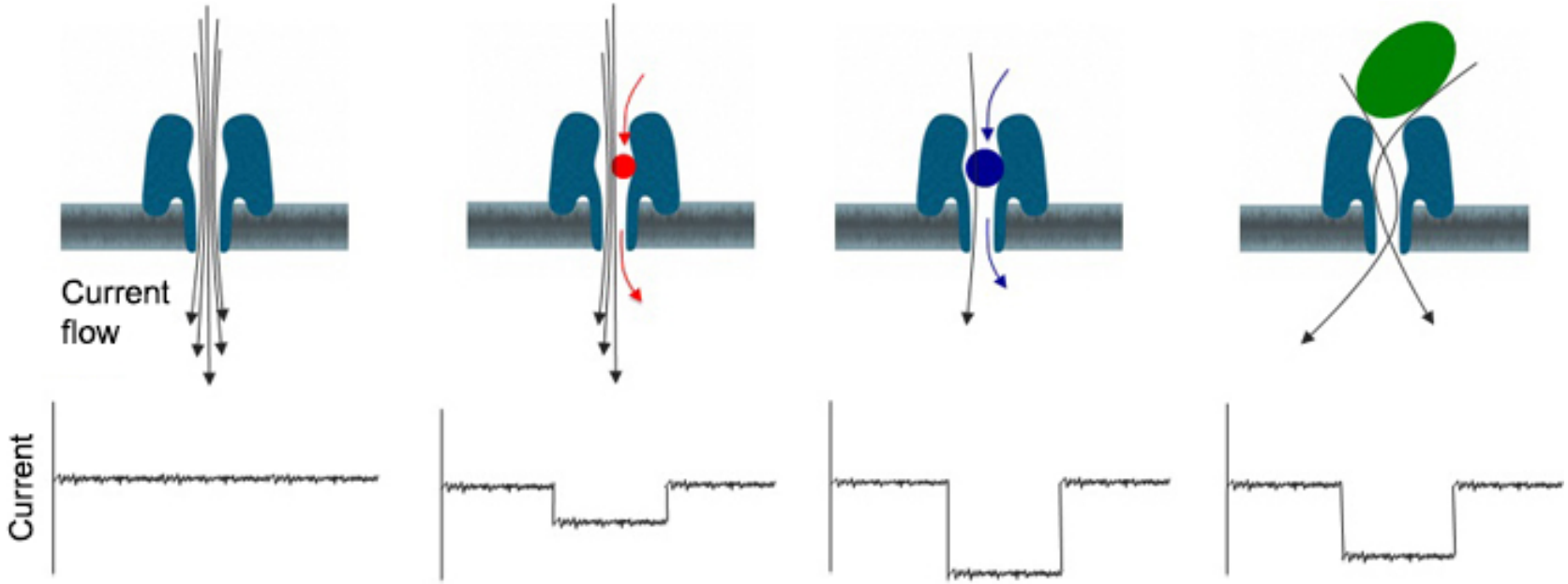
Nanopores DNA data. New sequencing hardware.

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.