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 $\pi$ such that

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

Recover $\pi$?

![Similarity matrix](image1)
![Input](image2)
![Reconstructed](image3)
Seriation has direct applications in \((de \ novo)\) genome assembly.

- Genomes are cloned multiple times and randomly cut into shorter reads (~ 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.
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 (~ 400bp), few errors (~ 2%). Repeats are challenging

■ Third generation: long reads (~ 10kbp), more errors (~ 15%). Can resolve repeats, but noise is challenging
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)
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 \( L_A = \text{diag}(A1) - A \) 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 \( x \in \mathbb{R}^n \),

\[
x^T L_A x = x^T \text{diag}(A1)x - x^T Ax
\]

\[
= \sum_{i=1}^{n} x_i^2(\sum_{j=1}^{n} A_{ij}) - \sum_{i,j=1}^{n} A_{ij}x_ix_j
\]

\[
= \sum_{i,j=1}^{n} A_{ij}(x_i^2 - x_ix_j)
\]

\[
= \frac{1}{2} \sum_{i,j=1}^{n} A_{ij}(x_i^2 + x_j^2 - 2x_ix_j)
\]

\[
= \frac{1}{2} \sum_{i,j=1}^{n} A_{ij}(x_i - x_j)^2
\]
Seriation and 2-SUM

Combinatorial Solution.

For certain matrices $A$, $\text{2-SUM} \iff \text{seriation}$. ([Fogel et al., 2013])
Spectral relaxation

2-SUM problem:

\[
\min_{\pi \in \mathcal{P}} \pi^T L_A \pi
\]

\textbf{NP-Complete} for generic matrices \( A \).

Set of permutation vectors:

\[ \pi_i \in \{1, \ldots, n\}, \quad \forall 1 \leq i \leq n \]
\[ \pi^T \mathbf{1} = \frac{n(n+1)}{2} \]
\[ \|\pi\|_2^2 = \frac{n(n+1)(2n+1)}{6} \]

Let \( c = \frac{n+1}{2} \mathbf{1} \). \( L_A \mathbf{1} = 0 \). Withdrawing \( c \) from any vector \( \pi \) 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

\[
f = \arg\min \ x^T L_A x.
\]
\[
\begin{align*}
1^T x &= 0, \\
\|x\|_2 &= 1
\end{align*}
\]
**Spectral Seriation.** Define the Laplacian of $A$ as $L_A = \text{diag}(A1) - A$, the Fiedler vector of $A$ is written

$$f = \arg\min_{x^T1=0, \|x\|_2=1} x^T L_A x.$$ 

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 $A \in S_n$ 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 $\Pi v$ 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

![Graph showing read length distribution with mean, median, min, max, and percentage of reads longer than 7Kbp](image)

A. Recanati

Institut Curie, Octobre 2016, 14/17
Application to genome assembly

Layout.

- Two bacterial genomes: *E. Coli* and *A. Baylyi*
- Circular genomes, size $\sim 4\text{Mbp}$
- 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

![Graph showing true ordering vs. recovered spectral ordering](image-url)
Conclusion

Straightforward assembly pipeline.

- Equivalence 2-SUM $\iff$ 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


Consensus

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

![Diagram of window layout and consensus](image)
Seriation

Combinatorial problems.

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

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

or equivalently

\[ \min_{\pi \in \mathcal{P}} \pi^T L_A \pi \]

where \( L_A \) is the Laplacian of \( A \).

- **NP-Complete** for generic matrices \( A \).
Convex Relaxation

Seriation as an optimization problem.

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

What’s the point?

- Gives a spectral (hence polynomial) solution for 2-SUM on some R-matrices.
- Write a \textbf{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

$$
\mathcal{D}_n = \{ X \in \mathbb{R}^{n \times n} : X \geq 0, X1 = 1, X^T1 = 1 \}
$$

is the convex hull of the set of permutation matrices.

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

- Solve

$$
\text{minimize} \quad \text{Tr}(Y^T \Pi^T L A \Pi Y) - \mu \| P \Pi \|_F^2
\quad \text{subject to} \quad e_1^T \Pi g + 1 \leq e_n^T \Pi g,
\Pi 1 = 1, \Pi^T 1 = 1,
\Pi \geq 0,
$$

in the variable $\Pi \in \mathbb{R}^{n \times n}$, where $P = I - \frac{1}{n} 11^T$ and $Y \in \mathbb{R}^{n \times p}$ is a matrix whose columns are small perturbations of $g = (1, \ldots, n)^T$. 

A. Recanati

Institut Curie, Octobre 2016, 22/17
Convex Relaxation

Objective. \[ \text{Tr}(Y^T \Pi^T L_A \Pi Y) - \mu \| P \Pi \|_F^2 \]

- **2-SUM** term \( \text{Tr}(Y^T \Pi^T L_A \Pi Y) = \sum_{i=1}^P y_i^T \Pi^T L_A \Pi y_i \) where \( y_i \) are small perturbations of the vector \( g = (1, \ldots, n)^T \).

- **Orthogonalization penalty** \(-\mu \| P \Pi \|_F^2\), where \( P = I - \frac{1}{n} \mathbf{1} \mathbf{1}^T \).
  - 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 \( \Pi \) doubly stochastic.
Convex Relaxation

Other relaxations.

- Relaxations for orthogonality constraints, e.g. SDPs in [Nemirovski, 2007, Coifman et al., 2008, So, 2011]. Simple idea: $Q^T Q = I$ is a quadratic constraint on $Q$, **lift it.** This yields a $O(\sqrt{n})$ approximation ratio.


- 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

\[ a \leq \pi(i) - \pi(j) \leq b \]

is written

\[ a \leq e_i^T \Pi g - e_j^T \Pi g \leq b. \]

which are linear constraints in \( \Pi \).

- **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 \( \sim 100 \) 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.