Convex Relaxations for Permutation Problems

Inria Junior Seminar

Fajwel Fogel

February 13, 2014

SIERRA team

My advisors: Alexandre d'Aspremont and Francis Bach

Convex optimization in 7 slides

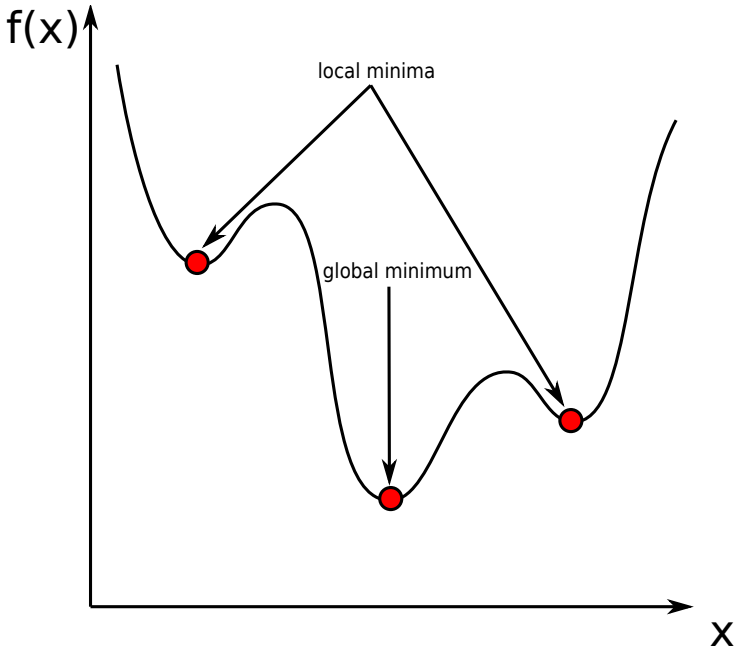
Application to DNA sequencing and archeology

Optimization is everywhere

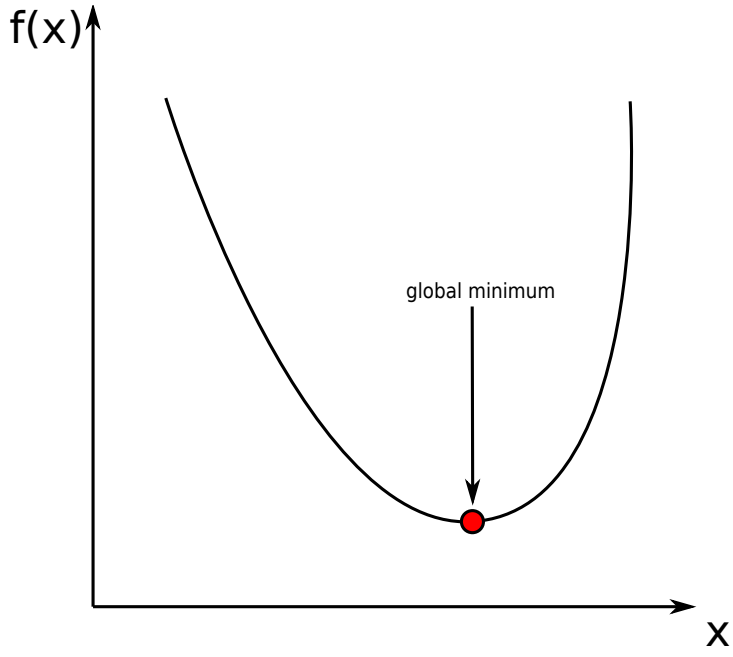
- Fit parameters of a model (statistics/machine learning, physics, bio-informatics...).
- Allocate ressources optimally (finance, transportation, operations research...).
- Any application when you think about it.
- Mathematically, an optimization problem is defined as minimizing a function f of a variable x subject to a set of constraints $x \in Q$.
- Of course x can be multidimensional.

Why convex?

- Optimization is everywhere, but most problems are very hard to solve!
- On the other hand: convex optimization problems can be solved globally and efficiently.
- Convex optimization problem: convex cost function and convex domain/constraints.



A non-convex function with local minima



A convex function with one global minimum

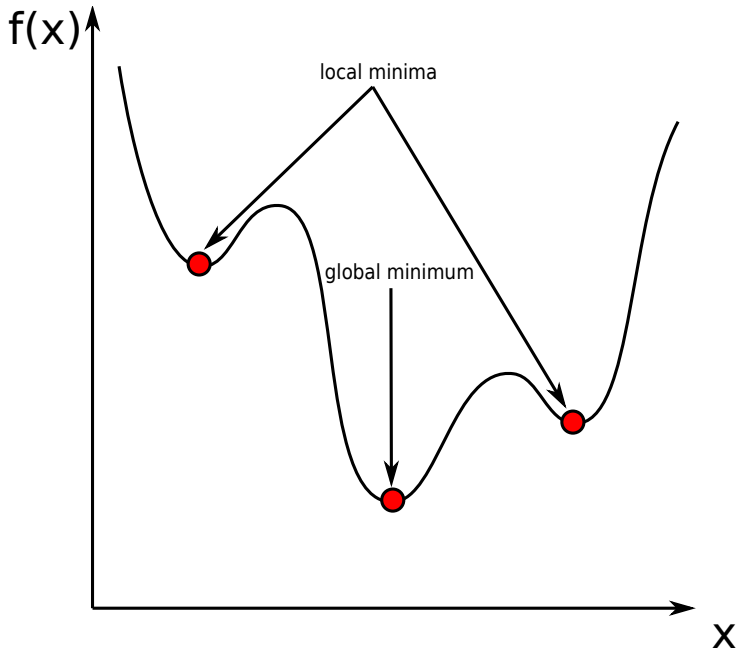
Convex optimization is a technology, for reasonable sized problems

- Many efficient and user-friendly solvers, including CVX, Mosek etc.
- Work very fast (micro second to few seconds) and give 10^{-12} accuracy solutions for problems of dimension < 1000.
- For bigger problems, use algorithms specifically tuned for the problem and parallelization when possible.

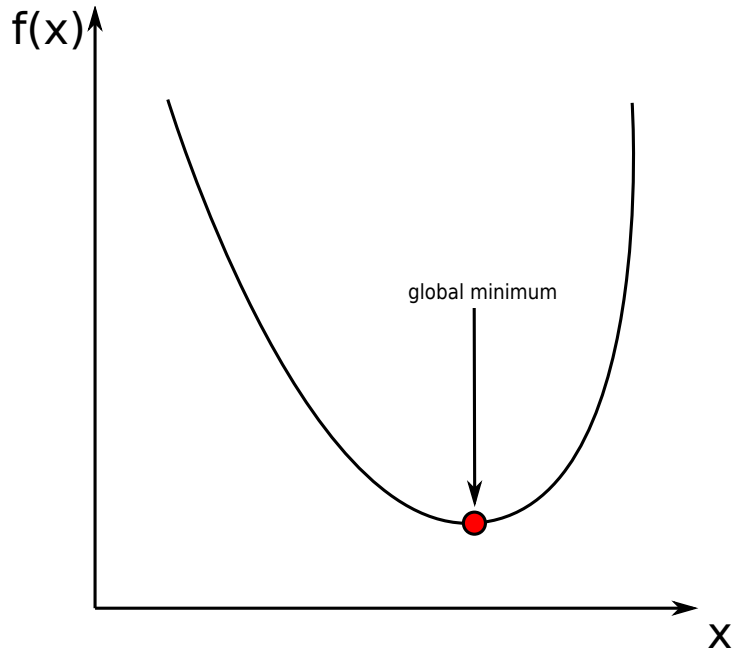
What about non convex problems?

Optimization is everywhere, but most problems are not convex!
Non convex optimization:

- finds local optima, with no guarantee on global optimality
- often relies on heuristics
- can work well in practice, but not in a systematic way.



A non-convex function with local minima



A convex function with one global minimum

Convex addict

- Can't we go back to the nice convex world?
- For some non-convex problems, it is possible to write a "relaxation" which gives an approximate solution to the original problem.
- When they work, relaxations can provide both good results and theoretical guarantees on hard problems.

How to relax a problem?

- Typical framework: convex objective function with non convex constraints.
- Relaxed problem: suppress non convex constraints/take the convex hull of the domain.

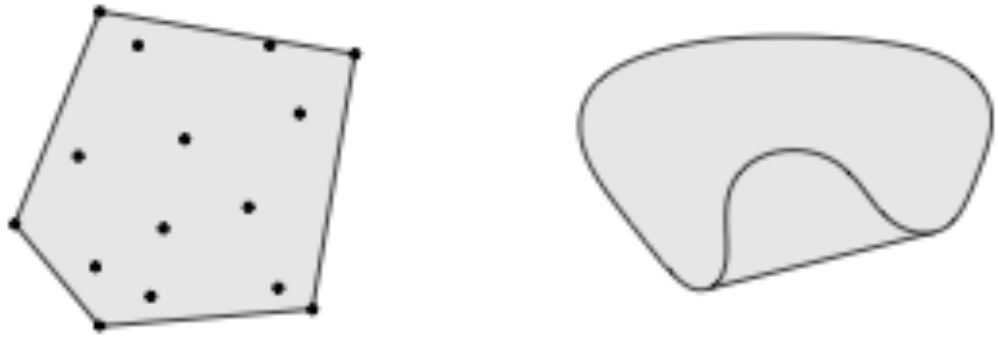


Figure 2.3 The convex hulls of two sets in $\mathbf{R}^2$. *Left.* The convex hull of a set of fifteen points (shown as dots) is the pentagon (shown shaded). *Right.* The convex hull of the kidney shaped set in figure 2.2 is the shaded set.

- Example: relax set of permutation matrices by set of doubly stochastic matrices (non-negative matrices whose rows and columns sum to one).

How to relax a problem?

- Project solution of relaxed problem on original set to get a feasible point.
- Get lower bound on the original problem: get an idea of how far you are from the true solution.

$$f(x^{\text{relax}}) \leq f(x^{\text{optimal}}) \leq f(x^{\text{projected}})$$

- For some problems it is possible to quantify the "tightness" of their relaxations.

Part two: a glimpse on my work.

Seriation

The Seriation Problem.

Randomly ordered movie.

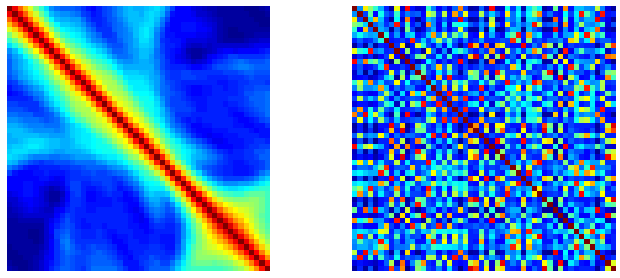


Image similarity matrix (true & observed)

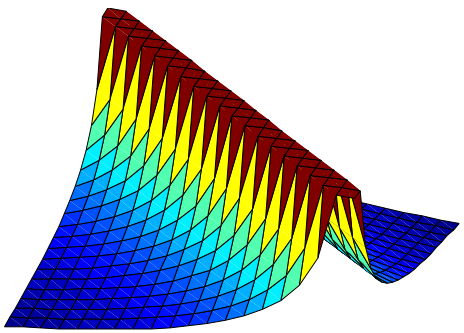
Reconstructed movie.

Seriation

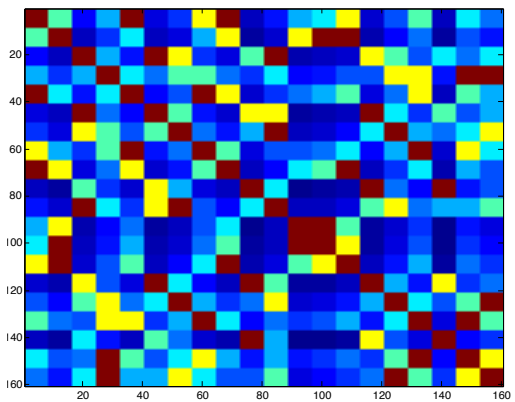
- Pairwise **similarity information** A_{ij} on n variables.
- Suppose the data has a **serial structure**, i.e. there is an order π such that

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

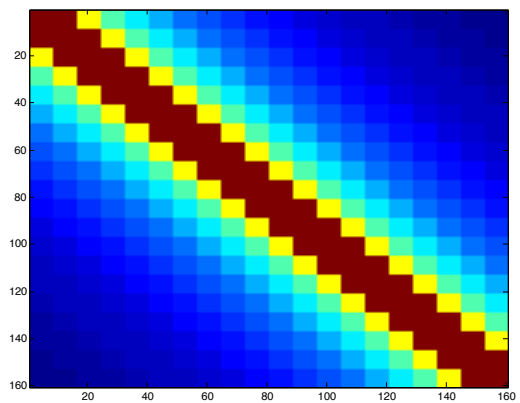
Recover π?



Similarity matrix



Input

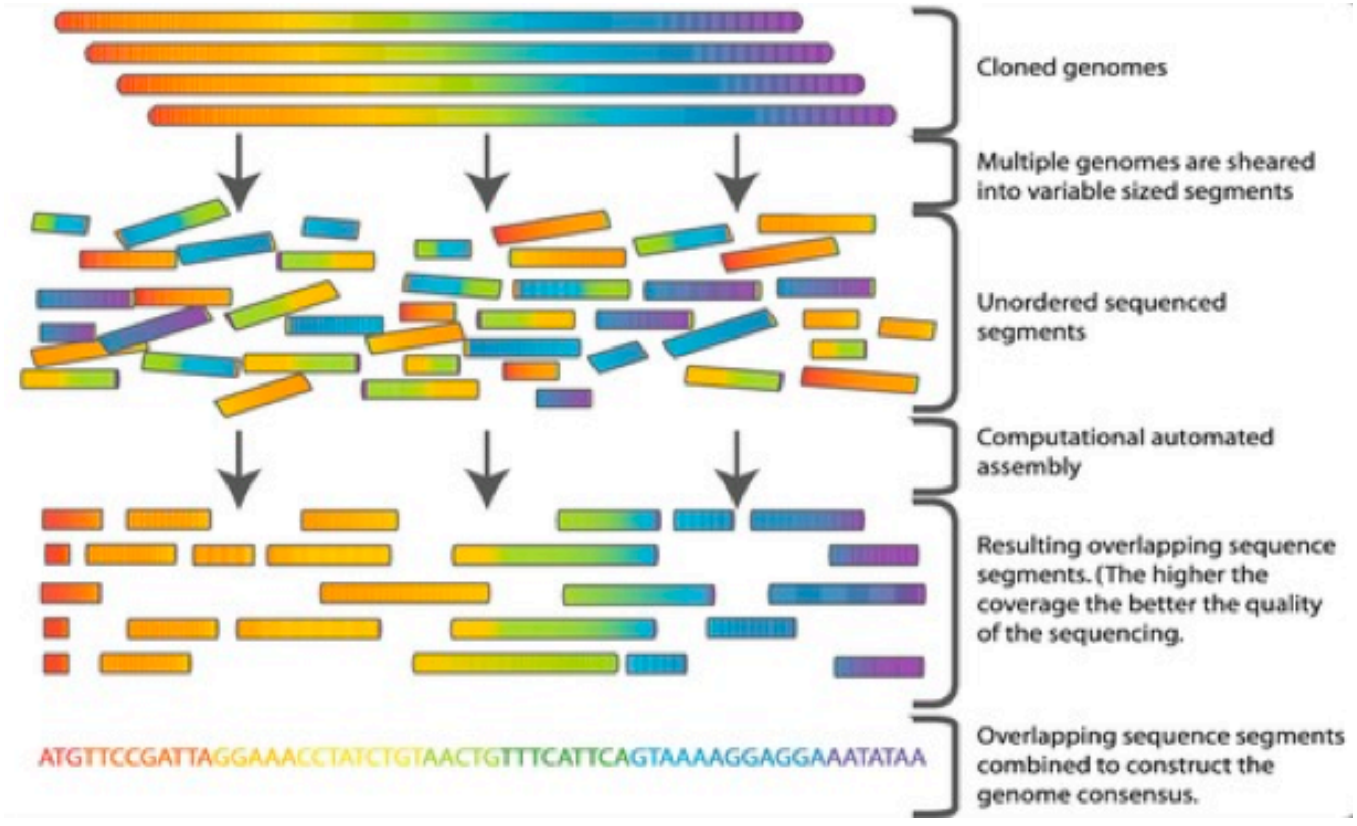


Reconstructed

Shotgun gene sequencing

C1P has direct applications in shotgun gene sequencing.

- Genomes are cloned multiple times and randomly cut into shorter reads
  ($\sim$ a few hundred base pairs), which are fully sequenced.
- Reorder the reads to recover the genome.



(from Wikipedia...)

Exact solution in the noiseless case

A "magical" result : the Fiedler vector reorders a R-matrix in the noiseless case!

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.

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.*

Convex relaxation

- Combinatorial objective:

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

  where L_A is the Laplacian of A and $y = (1, \ldots, n)^T$.
- Π permutation matrix if and only Π is both **doubly stochastic** and **orthogonal.**
- Set of doubly stochastic matrices is convex hull of permutation matrices
- Relax set of permutations by removing orthogonality constraint: [Fogel, Jenatton, Bach, and d'Aspremont, 2013]

$$\begin{array}{ll} \text{minimize} & y^T \Pi^T L_A \Pi y \\ \text{subject to} & e_1^T \Pi y + 1 \leq e_n^T \Pi y, \\ & \Pi \mathbf{1} = \mathbf{1}, \, \Pi^T \mathbf{1} = \mathbf{1} \\ & \Pi \geq 0, \end{array}$$

in the variable $\Pi \in \mathbb{R}^{n \times n}$.

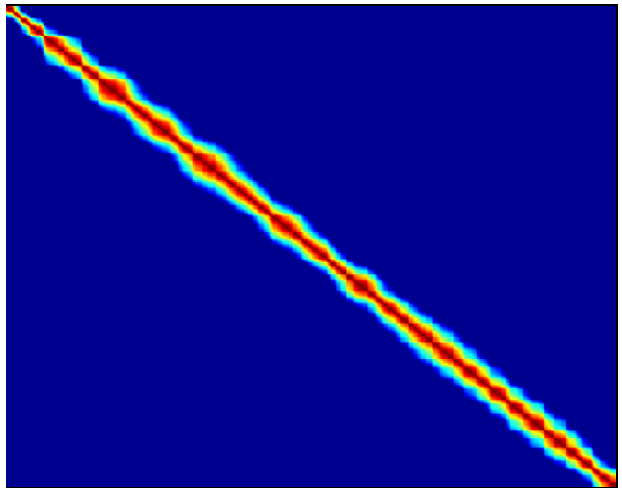
Convex relaxation

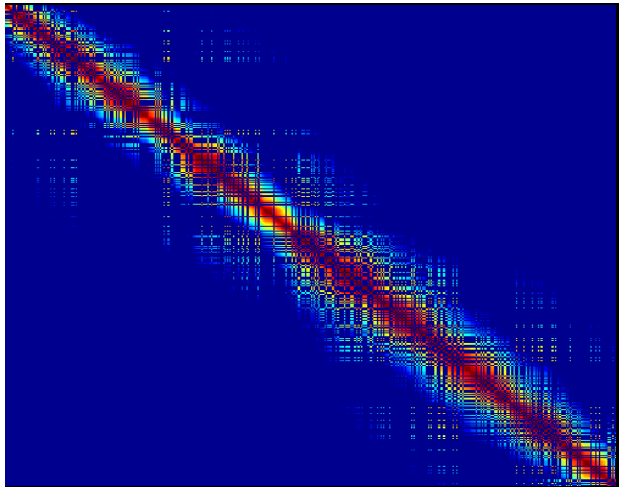
- Actually need a little more to make it work.
- Can add a priori information on the order we want to recover (e.g. we know that element i should be at distance d from element j).
- More robust to noise than spectral seriation, but not exact in noiseless case.
- Not yet scalable to datasets > 1000 points, but can use spectral seriation first and then refine with convex relaxation.

Numerical experiments

DNA

Reorder the *read* similarity matrix to solve C1P on 250 000 reads from human chromosome 22.

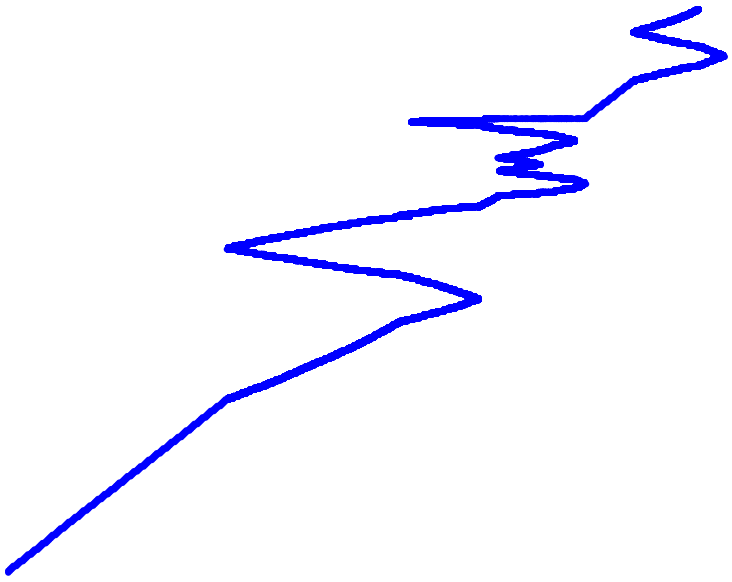




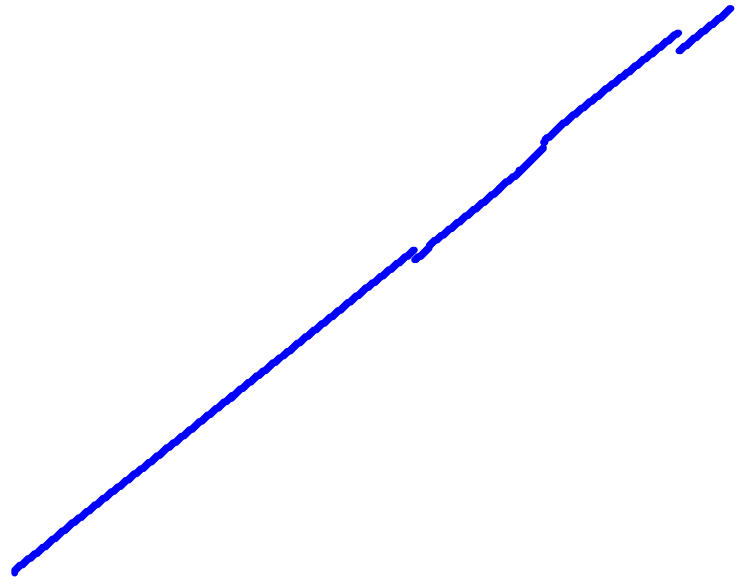
$\#$ *reads* $\times$ $\#$ *reads* matrix measuring the number of common k-mers between read pairs, reordered according to the spectral ordering.
The matrix is 250 000 $\times$ 250 000, we zoom in on two regions.

DNA

250 000 reads from **human chromosome 22.**



Spectral



Spectral + QP

Recovered read position versus true read position for the **spectral solution** and the **spectral solution followed by semi-supervised seriation**.
We see that the number of misplaced reads significantly decreases in the semi-supervised seriation solution.

Dead people

Row ordering, **70 artifacts** $\times$ **59 graves** matrix [Kendall, 1971]. Find the chronology of the 59 graves by making artifact occurrences contiguous in columns.

Kendall

Spectral

QP

The **Hodson's Munsingen dataset:** column ordering given by Kendall *(left)*, Fiedler solution *(center)*, best unsupervised QP solution from 100 experiments with different Y, based on combinatorial objective *(right)*.

Merci de votre attention!

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.

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

M. Gilchrist. Bringing it all back home: Next-generation sequencing technology and you. In *Mill Hill Essays 2010*. 2010.

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

David G Kendall. Abundance matrices and seriation in archaeology. *Probability Theory and Related Fields*, 17(2): 104–112, 1971.