Omic Fold Changes Clustering With Alignment & Network Inference

A statistical approach to study the radiation response of endothelial cells

POLINA ARSENTEVA[1,2] & VINCENT PAGET[2] & OLIVIER GUIPAUD[2] & FABIEN MILLIAT[2] & HERVÉ CARDOT[1] & MOHAMED AMINE BENADJAOUD[3]

[1] IMB, Université de Bourgogne [2] IRSN, PSE-SANTE/SERAMED/LRMed [3] IRSN, PSE-SANTE/SERAMED

I. Introduction

Context: treatment of cancer by radiotherapy.

- **Problem:** undesirable effects for the healthy tissues situated in close proximity to the irradiated tumors.
- **Ultimate goal:** compare different radiation treatment configurations in order to suggest a treatment associated with minimal risk.

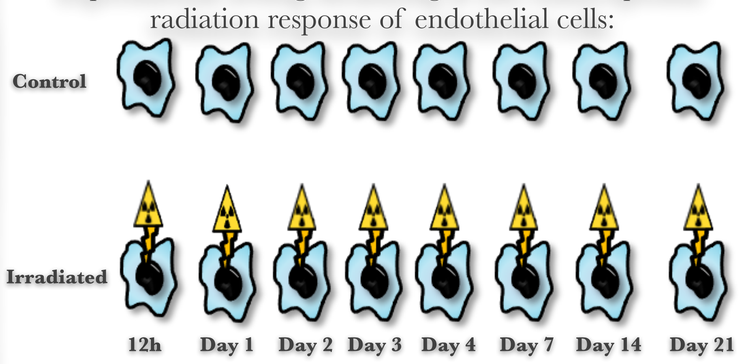
Focus of this work: the dynamic of endothelial cells' response to irradiation.

- **Why endothelium?** Key cell compartment for the healthy tissue radiation response and the occurrence of side effects.

Goal:	Tool:
Extract key features of data describing response to different types of irradiation	Clustering & Alignment
Propose powerful visualisation tools for biological interpretation of the results	Network Inference

II. From data to radio-induced fold changes

Experimental setting measuring *in-vitro* transcriptomic radiation response of endothelial cells:



Challenges:

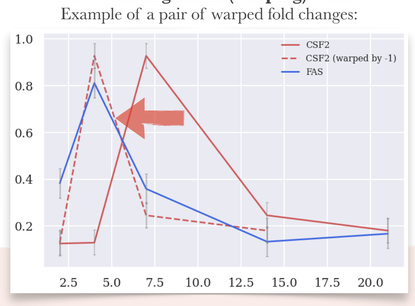
- Independent (destructive sampling) & unequally spaced time points
  ➡ Not time series data
- Small number of time points
  ➡ Functional data approach not applicable
- ~100-500 observed genes
  ➡ Correlations & computational cost to consider
- Multiple replicates
  ➡ Uncertainties to consider

Response: Y^t_{ikj}

- Gene $i \in \{1,2,...,n_e\}$
- Time point $t \in \{t_1, t_2, \dots, t_p\}$
- Experimental condition: $k = \begin{cases} 0 \text{ if control} \\ 1 \text{ if irradiated} \end{cases}$
- Replicate $j \in \{1,2,...,n_r\}$

$$\text{Fold Change}(\text{time}) = \overline{\text{Effect}}_{\text{irradiated}}(\text{time}) - \overline{\text{Effect}}_{\text{control}}(\text{time})$$

How does alignment (warping) work?
Example of a pair of warped fold changes:



III. Clustering & Alignment

- Fold changes' estimators as **random variables**: for a pair of genes i and i',

$$\Sigma_{\Gamma_i} = \begin{bmatrix} \sigma^2_{\Gamma_i^{t_1}} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{\Gamma_i^{t_p}} \end{bmatrix} \text{ where } \sigma^2_{\Gamma_i^t} = \frac{\sum_{j=1}^{n_r}\left[(Y^t_{i1j} - \overline{Y^t_{i1}})^2 + (Y^t_{i0j} - \overline{Y^t_{i0}})^2\right]}{n_r - 1}$$

$$\Gamma_i = \left(\overline{Y^{t_1}_{i1}} - \overline{Y^{t_1}_{i0}}, \dots, \overline{Y^{t_p}_{i1}} - \overline{Y^{t_p}_{i0}}\right)$$

Joint distribution:

$$\widehat{\Gamma_{ii'}} \sim \mathcal{N}\left(\begin{bmatrix}\Gamma_i \\ \Gamma_{i'}\end{bmatrix}, \begin{bmatrix}\Sigma_{\Gamma_i} & K \\ K^T & \Sigma_{\Gamma_{i'}}\end{bmatrix}\right)$$

$$K = \begin{bmatrix} \rho_{\Gamma_i\Gamma_{i'}^{t_1}} & & 0 \\ & \ddots & \\ 0 & & \rho_{\Gamma_i\Gamma_{i'}^{t_p}} \end{bmatrix} \text{ where } \rho_{\Gamma_i\Gamma_{i'}^t} = \frac{\sum_{j=1}^{n_r}\left[(Y^t_{i1j} - \overline{Y^t_{i1}})(Y^t_{i'1j} - \overline{Y^t_{i'1}}) + (Y^t_{i0j} - \overline{Y^t_{i0}})(Y^t_{i'0j} - \overline{Y^t_{i'0}})\right]}{n_r - 1}$$

- **Distance** $\widehat{\mathbf{d}_2^2}$: L^2-distance between normal r. v. with estimated joint distribution:

$$\widehat{\mathbf{d}_2^2}\left(\widehat{\Gamma}_i, \widehat{\Gamma}_{i'}\right) = \sum_{l=1}^{p}\left(\Gamma_i^{t_l} - \Gamma_{i'}^{t_l}\right)^2 + \sum_{l=1}^{p}\sigma^2_{\Gamma_i^{t_l}} + \sum_{l=1}^{p}\sigma^2_{\Gamma_{i'}^{t_l}} - 2\sum_{l=1}^{p}\rho_{\Gamma_i\Gamma_{i'}^{t_l}}$$

- **Base algorithm:** k-medoids initiated with k-means++.
- **Integrating alignment:** clustering of warped fold changes pairs

$$\begin{bmatrix}\widehat{\Gamma_i \circ \mathscr{W}_s} \\ \widehat{\Gamma_{i'} \circ \mathscr{W}_s}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\Gamma_i \circ \mathscr{W}_s \\ \Gamma_{i'} \circ \mathscr{W}_s\end{bmatrix}, \begin{bmatrix}\Sigma_{\Gamma_i} \circ \mathscr{W}_s & K \circ \mathscr{W}_s \\ (K \circ \mathscr{W}_s)^\intercal & \Sigma_{\Gamma_{i'}} \circ \mathscr{W}_s\end{bmatrix}\right)$$

Time warp aligning a pair of time vectors by step s: $\mathscr{W}_s : \mathscr{T}^2 \to \mathscr{T}^2$, $\begin{pmatrix}t^0 \\ t^0\end{pmatrix} \mapsto \begin{pmatrix}t^{11} \\ t^{12}\end{pmatrix}$

- **Replacing** $\widehat{\mathbf{d}_2^2}$ by a more general distance $\widehat{\mathbf{diss}}$:

$$\widehat{\mathbf{diss}}\left(\widehat{\Gamma_i \circ \mathscr{W}_s}, \widehat{\Gamma_{i'} \circ \mathscr{W}_s}\right) = \|\Gamma_i \circ \mathscr{W}_s - \Gamma_{i'} \circ \mathscr{W}_s\|^2 + \sum(\Sigma_{\Gamma_i} \circ \mathscr{W}_s) + \sum(\Sigma_{\Gamma_{i'}} \circ \mathscr{W}_s) - 2\sum(K \circ \mathscr{W}_s)$$

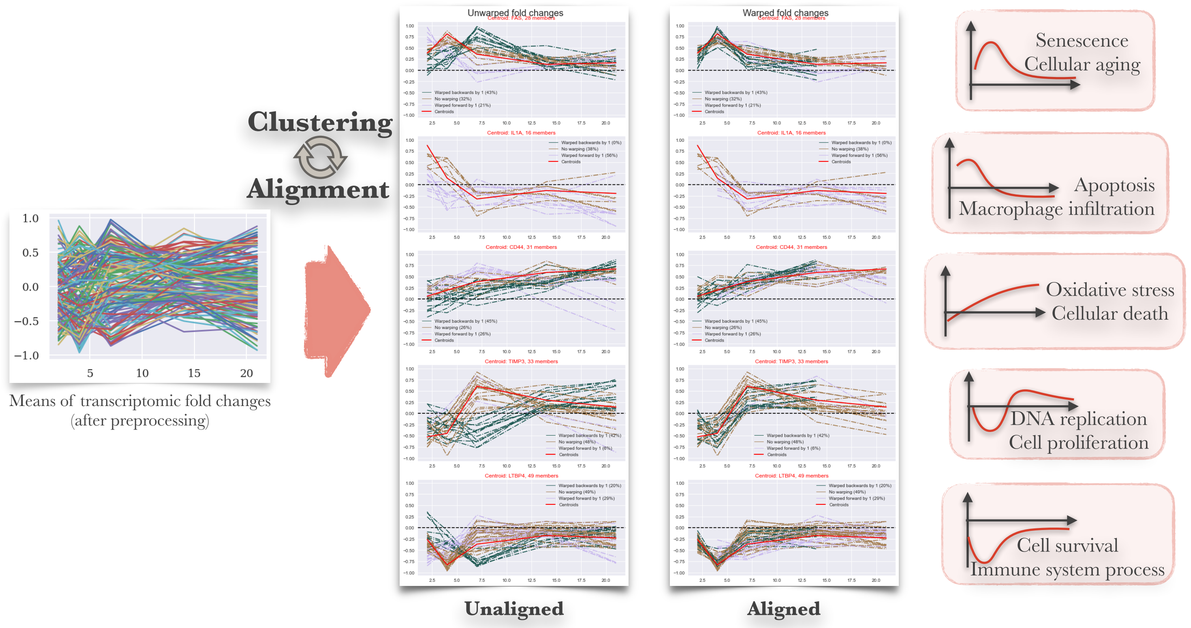
- Clustering based on the **Optimal Warping Distance** matrix:

$$\mathscr{OWD} = \left[\min_{s \in \mathcal{S}}\left[\widehat{\mathbf{diss}}\left(\widehat{\Gamma_i \circ \mathscr{W}_s}, \widehat{\Gamma_{i'} \circ \mathscr{W}_s}\right)\right]\right]_{1 \le i,i' \le n_e}$$

➡ Clustering & Alignment performed simultaneously with a low computational cost.

IV. Clustering & Alignment of real fold changes

Applied to a real transcriptomic dataset, the procedure allowed to identify **5 distinct gene response types**. Enrichment analysis with Pathway Studio shows the existence of associated distinct **cellular functions**:



V. Network Inference

- **Similarity** measure based on the post-warping distances:

$$Sim\left(\widehat{\Gamma}_i, \widehat{\Gamma}_{i'}\right) = \frac{\max_{(a,b)\in\{1,\dots,n_e\}^2} \mathscr{OWD}_{ab} - \mathscr{OWD}_{ii'}}{\max_{(a,b)\in\{1,\dots,n_e\}^2} \mathscr{OWD}_{ab}}$$

- Binary **adjacency matrix** $X = (X_{ii'})_{(i,i')\in\{1,\dots,n_e\}^2}$ describing directed graph:

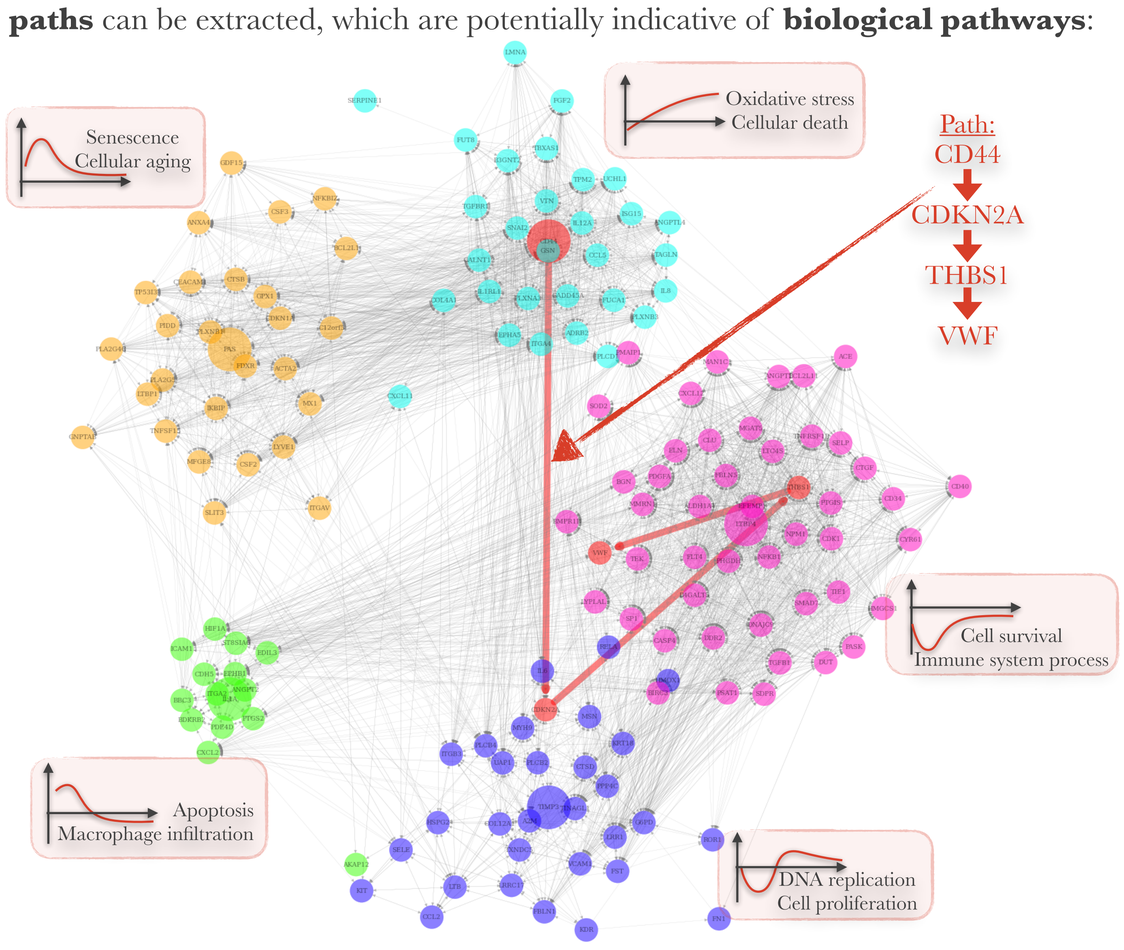
$$X_{ii'} = \mathbb{1}_{Sim\left(\widehat{\Gamma}_i, \widehat{\Gamma}_{i'}\right) \ge \mathbf{q}} \times \mathbb{1}_{\mathscr{OW}_{ii'} \ge 0}$$

Edge $i \to i'$ exists if i and i' are at least as similar is the thresholding parameter $\mathbf{q}$

Edge $i \to i'$ exists if the optimal warp is non-negative

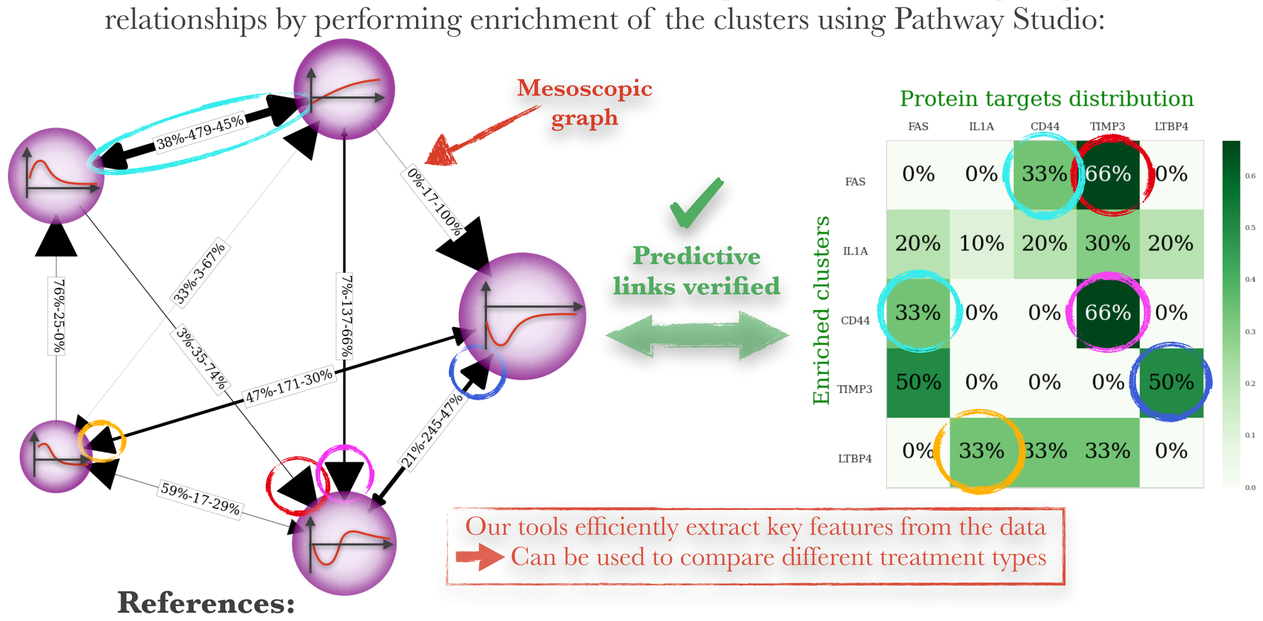
Optimal Warp matrix

Network is visualised in a **block form**, where blocks are the inferred clusters. Gene **paths** can be extracted, which are potentially indicative of **biological pathways**:



VI. Our approach vs. Biological literature

We propose a **mesoscopic** representation of the network summarising the key characteristics extracted from the data. We compare it to the known gene-protein relationships by performing enrichment of the clusters using Pathway Studio:



Our tools efficiently extract key features from the data
➡ Can be used to compare different treatment types

References:

1. Clark R. Givens, Rae Michael Shortt. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal.*
2. David Arthur, Sergei Vassilvitskii. (2007). k-means++: the advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, SIAM, pp. 1027-1035.
3. Hae-Sang Park, Chi-Hyuck Jun. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Syst. Appl.*. 36. 3336-3341.
4. Pathway Studio: https://www.pathwaystudio.com

All data were produced by the IRSN and are not public yet. Code & data soon to be available at: github.com/parsenteva.

Conflict of interest disclosure: no relevant relationships with any commercial or non-profit organisations.