Omic fold changes clustering with alignment and network inference: a statistical approach to study the radiation response of endothelial cells

Polina Arsenteva [1] & Vincent Paget [2] & Olivier Guipaud [3] & Fabien Milliat [4] & Hervé Cardot [5] & Mohamed Amine Benadjaoud [6]

[1,5] *IMB, Université de Bourgogne. 9 avenue Alain Savary, 21078 Dijon Cedex*
[2,3,4] *IRSN, PSE-SANTE/SERAMED/LRMed.*
[6] *IRSN, PSE-SANTE/SERAMED.*
31 avenue de la Division Leclerc, 92262 Fontenay-aux-Roses Cedex
[1] *polina.arsenteva@u-bourgogne.fr* [2] *vincent.paget@irsn.fr*
[3] *olivier.guipaud@irsn.fr* [4] *fabien.milliat@irsn.fr*
[5] *herve.cardot@u-bourgogne.fr* [6] *mohamedamine.benadjaoud@irsn.fr*

Résumé. En radiobiologie, un "fold change" est défini comme une mesure de différence entre l'état irradié et non-irradié d'une certaine entité biologique au cours du temps. L'objectif de ce travail est de déterminer un petit nombre des types de comportements les plus représentatifs parmi les nombreuses entités biologiques considérées, ainsi que d'identifier les pathways biologiques potentiels liés à la réponse à la radiothérapie. Nous proposons une procédure qui consiste à effectuer simultanément un clustering et un alignement des estimateurs temporels des fold changes suivis d'une inférence de réseau en utilisant un modèle à blocs stochastiques. Le clustering est effectué par rapport aux lois de probabilité, ce qui est motivé par la nécessité de tenir compte des incertitudes et des corrélations entre les variables. L'approche proposée a été évaluée à travers de nombreuses études de simulation et montre des perspectives prometteuses sur des données réelles.

Mots-clés. Irradiation in vitro, données omiques, clustering de lois de probabilité, time warping, modèle à blocs stochastiques.

Abstract. In radiobiology, a fold change is defined as a measure of difference between irradiated and non-irradiated condition for a certain biological entity over time. The goal of this work is to determine a small number of the most representative behavior types among the numerous biological entities that are considered, as well as to identify potential biological pathways linked to the response to radiotherapy. We propose a procedure that consists in performing simultaneous clustering and alignment of fold changes' temporal estimators followed by network inference using stochastic block model. The clustering is distribution-based, which is motivated by the need to account for uncertainties and correlations between variables. The proposed approach has been evaluated through numerous simulation studies and show promising prospects on real data.

Keywords. In vitro irradiation, omic data, distribution clustering, time warping, stochastic block model.

1 Context

More than 200 000 patients undergo radiotherapy in France every year. Similarly to other treatments, it may induce adverse side effects for healthy tissues situated close to the irradiated tumor. It is thus of substantial importance to study and compare different modes of radiotherapy that vary in dose, volume, energy, etc. with a goal of selecting such that minimize the potential undesirable consequences.

This work focuses on studying cellular response to irradiation, namely that of endothelial cells, a key actor in the appearance of radiation adverse effects. Specifically, we study the expression of different biological entities originating from multiple omics in vitro datasets (e.g. transcriptomic measuring gene expression, proteomic for protein expression) that were collected for several time points. The common feature of all datasets is the presence of two experimental conditions: irradiated and non-irradiated. The quantity of interest is radio-induced fold change: a measure of irradiation effect represented by the difference between the two experimental conditions over time.

2 Methodology

We consider an observation Y^t_{ikj} from one of the studied datasets such that $i \in \{1, 2, \dots, n_e\}$ denotes an entity (e.g. a gene in case of transcriptomic data) where n_e is the number of considered biological entities, $j \in \{1, 2, \dots, n_r\}$ indicates a replicate, k is the experimental condition such that $k = 0$ if control and $k = 1$ if irradiated, and $t \in \{t_1, t_2, \dots, t_p\}$ stands for a time point. The presence of multiple replicates in the datasets allows to estimate the joint distributions of all variable pairs representing a fold change while taking into account the correlations between genes, i.e. it is possible that $\mathrm{Cov}\left(Y^t_{ikj}, Y^t_{i'kj}\right) \neq 0$ for $i \neq i'$. In the course of this project, multivariate ANOVA model is used to estimate fold changes as vectors with respect to time after irradiation, obtaining deterministic estimates of individual fold changes denoted $\Gamma_i = \left(\Gamma_i^{t_1}, \dots, \Gamma_i^{t_p}\right)$. Since these estimations are subject to uncertainties that should be considered in subsequent investigations, we consider the estimators of the fold changes as random Gaussian vectors:

$$\widehat{\Gamma}_i \sim \mathcal{N}\left(\Gamma_i, \Sigma_{\Gamma_i}\right) \text{ such that } \Sigma_{\Gamma_i} = \begin{bmatrix} \sigma^2_{\Gamma_i^{t_1}} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{\Gamma_i^{t_p}} \end{bmatrix}$$

$$\text{and } \sigma^2_{\Gamma_i^t} = \frac{\sum_{j=1}^{n_r} \left[(Y^t_{i1j} - \overline{Y^t_{i1}})^2 + (Y^t_{i0j} - \overline{Y^t_{i0}})^2\right]}{n_r - 1}.$$

A diagonal covariance matrix is considered since we do not have access to timewise covariances due to the fact that the measures for different time points are produced individually,

hence they are not correlated, i.e. $\text{Cov}\left(Y_{ikj}^{t}, Y_{i'kj}^{t'}\right) = 0$ for $(i, i') \in \{1, 2, \dots, n_e\}^2$ and $t \neq t'$.

As a result, we obtain random estimators of pairs of fold changes' $\widehat{\Gamma}_i$ and $\widehat{\Gamma}_{i'}$ for $(i, i') \in \{1, 2, \dots, n_e\}^2$ with the joint distribution $\begin{bmatrix} \widehat{\Gamma}_i \\ \widehat{\Gamma}_{i'} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \Gamma_i \\ \Gamma_{i'} \end{bmatrix}, \begin{bmatrix} \Sigma_{\Gamma_i} & K \\ K^\intercal & \Sigma_{\Gamma_{i'}} \end{bmatrix}\right)$. In order to perform the clustering of the fold changes, an appropriate distance between distributions needs to be chosen. We introduce a distance $\widehat{\mathbf{d_2^2}}$, derived from L^2-distance between normally distributed fold changes estimators [1], i.e:

$$\widehat{\mathbf{d_2^2}}\left(\widehat{\Gamma}_i, \widehat{\Gamma}_{i'}\right) = \mathbb{E}\|\widehat{\Gamma}_i - \widehat{\Gamma}_{i'}\|^2 = \|\Gamma_i - \Gamma_{i'}\|^2 + \text{Tr}(\Sigma_{\Gamma_i}) + \text{Tr}(\Sigma_{\Gamma_{i'}}) - 2\text{Tr}(K)$$

using K estimated from the data. This choice allows to fully take the correlations between the entities into account during clustering. As for the choice of clustering algorithm, k-medoids has been selected as an alternative to k-means that allows to keep track of all the correlations [2][3]. This clustering procedure is performed combined with time warping, it is computationally efficient and is able to capture various relationships between entities.

The final step in the procedure corresponds to omic fold changes' network inference with stochastic block model [4]. The network is defined by the adjacency matrix constructed using distances between the fold changes. When inferring communities, we use the initialization sensitivity of stochastic block model by initializing the parameters of Variational Expectation Maximization (VEM) algorithm with the previously obtained clusters. As a result, we slightly improve the clusters that remain similar to those obtained with $\widehat{\mathbf{d_2^2}}$-based k-medoids on the one hand, and establish a comprehensive link between the network and the clusters through a model-based approach on the other hand.

3 Results and perspectives

In order to validate the approach, multiple simulation studies have been conducted. The simulated data attempt to mimic as closely as possible the real fold changes and their characteristics that are expected to be captured by the proposed procedure. Different simulation scenarios were used to study different aspects of the approach, such as the role of correlations between entities and the effect of time warping. In particular, the choices of $\widehat{\mathbf{d_2^2}}$ for the distance and k-medoids for the clustering algorithm were validated through comparison with Wasserstein and Hellinger distances, and k-means and hierarchical clustering algorithms respectively. It has been shown that our approach is the most adapted for the clustering of random fold changes.

First results of application to the real data have already been obtained. At this point a small number of distinct response types can be distinguished, with multiple temporal shifts within each group due to time warping. Currently, the final stage of analysis performed by

the biologists is taking place, which will allow to validate the approach through comparison with the existing knowledge in relation with the response to irradiation.

References

[1] Clark R. Givens, Rae Michael Shortt. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*.

[2] Hae-Sang Park, Jong-Seok Lee, Chi-Hyuck Jun. (2006). A K-means-like Algorithm for K-medoids Clustering and Its Performance. *Expert Syst Appl*, 36(2):3336–3341, DOI 10.1016/j.eswa.2008.01.039

[3] Erich Schubert, Peter J. Rousseeuw. (2019). Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *Similarity Search and Applications*, Springer International Publishing, 11807.

[4] Clement Lee, Darren J. Wilkinson. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science* 4, 122 (2019). DOI 10.1007/s41109-019-0232-2