

Machine Learning en Biologie des systèmes

Med Amine BENADJAUD

**PSE-Santé
Service de radiobiologie
et de médecine
régénérative (SERAMED)**

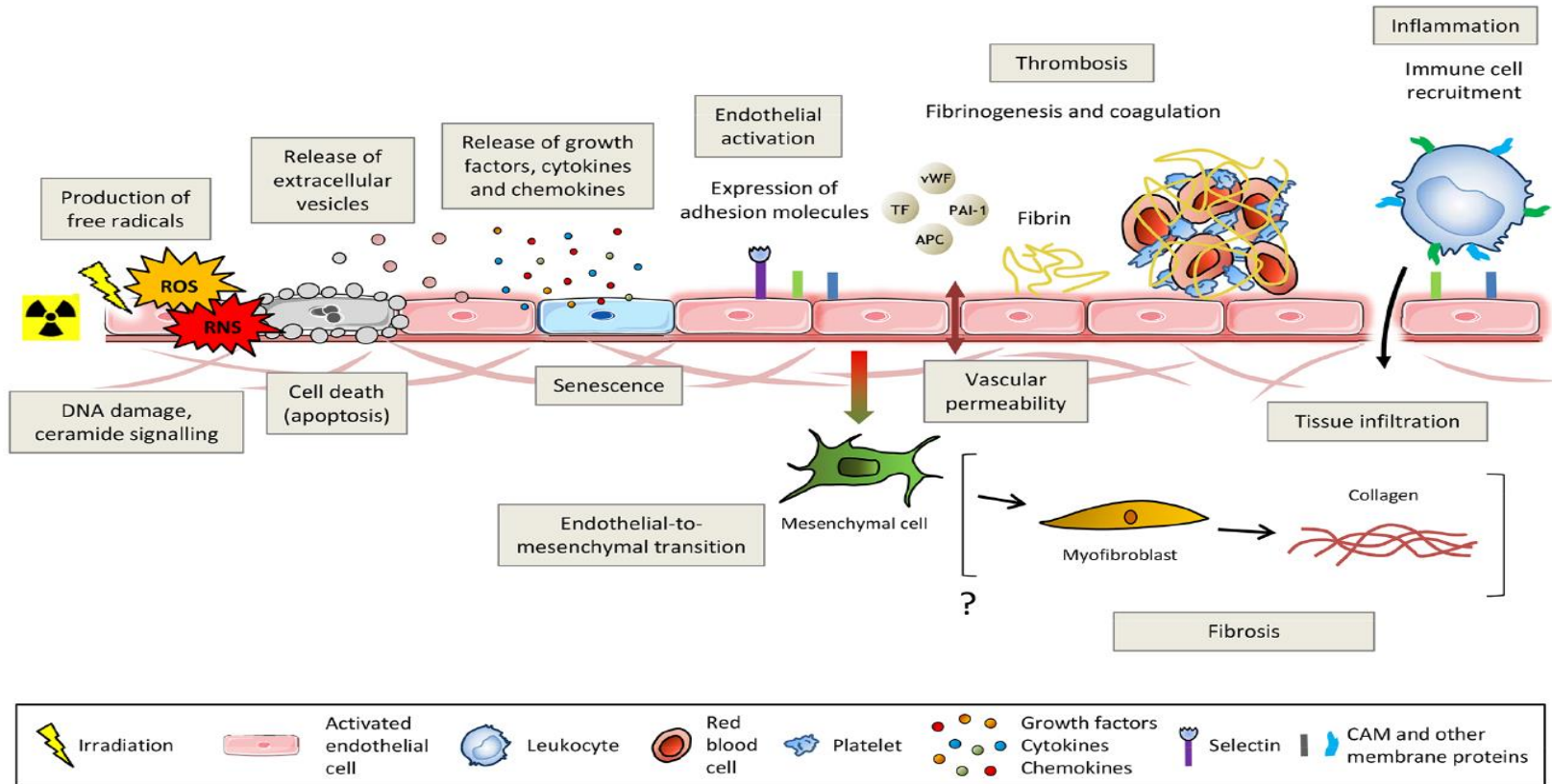
mohamedamine.benadjaoud@irsn.fr

Journée ECCOREV « intelligence artificielle et big data »

CEREG, domaine de l'Arbois, Aix-en-Provence

Contexte

- En France, **200 000 patients** traités chaque année par radiothérapie.
- 5 à 10 % des patients traités développent **des effets secondaires** tels que la **fibrose radio-induite** (*accumulation excessive de matrice extracellulaire dans un tissu*).



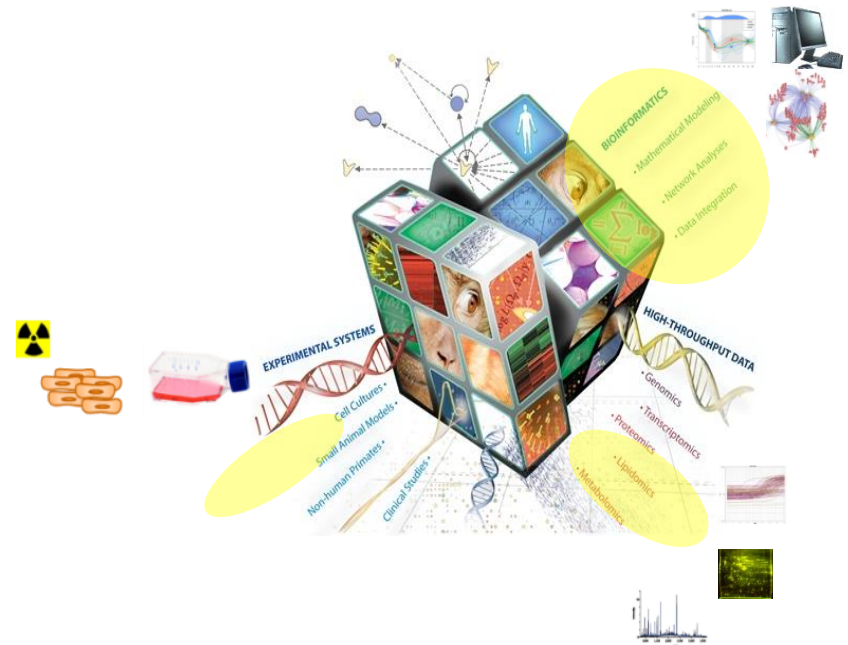
Endothélium vasculaire : couche de cellules qui tapissent la paroi interne des vaisseaux sanguins

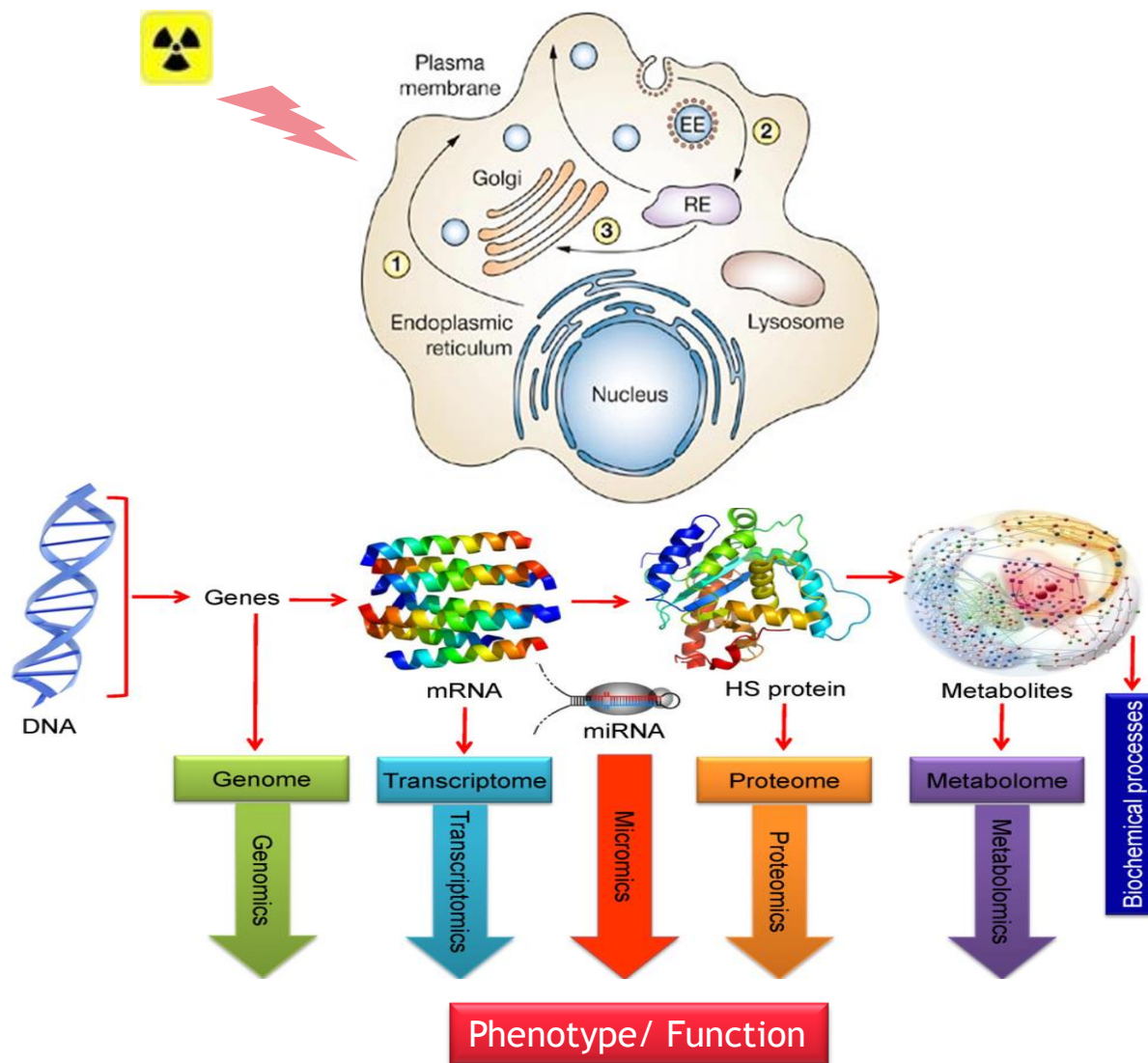
The importance of the vascular endothelial barrier in the immune-inflammatory response induced by radiotherapy

O Guipaud, C Jaillet, K Clément-Colmou, A François, S Supiot, and F Milliat (British Journal of Radiobiology 2018 Apr 20:20170762)

Introduction: La radiobiologie

- Une approche de **biologie des systèmes** pour comprendre les événements moléculaires impliqués dans la dysfonction **radio-induite** des cellules **endothéliales**.
- **Approche Holistique** visant à comprendre dans sa globalité les comportements **des réseaux moléculaires**, et en particulier leurs aspects **dynamiques**, ce qui requiert de construire des modèles **mathématiques** à partir des données expérimentales.
- **Expertise multidisciplinaire**
Radiobiologie, biologie moléculaire, biochimie, « omiques », bioinformatique, mathématiques





La réponse cellulaire est **intégrée, dynamique (espace et temps) et multi-échelle**

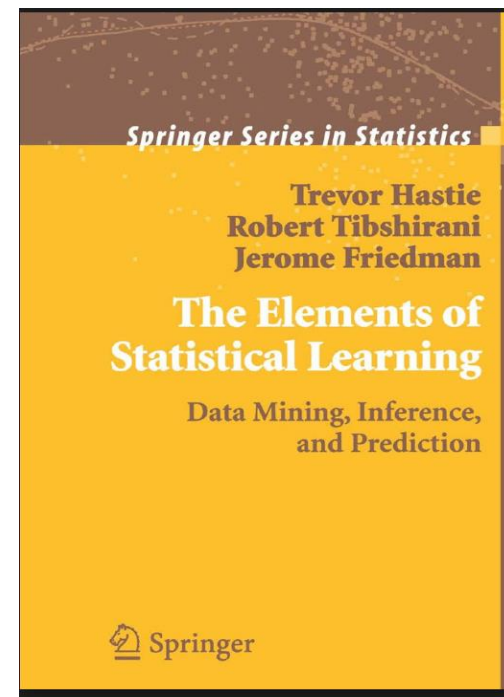
Machine Learning

Qu'est ce que le Machine Learning ?

“A machine learning algorithm is an algorithm that is able to learn from data ”

What do we mean by learning?

To extract **important patterns** and trends, and understand “what the data says.” We call this learning from data.



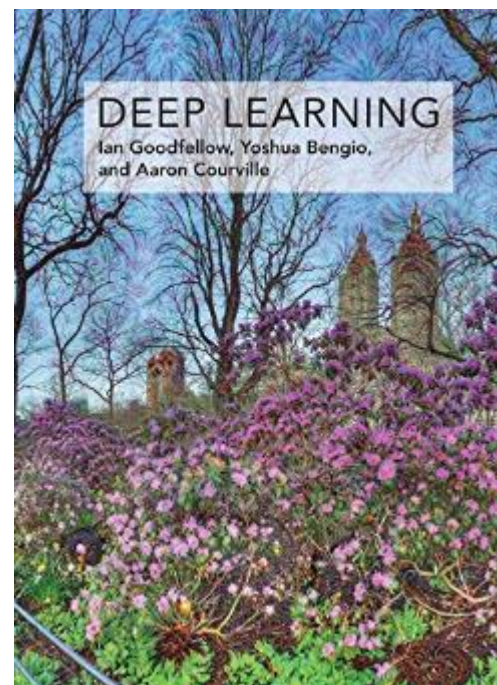
Machine Learning

Qu'est ce que le Machine Learning ?

“A machine learning algorithm is an algorithm that is able to learn from data ”

What do we mean by learning?

“A computer program is said to learn from **experience E** with respect to some class of **tasks T** and performance **measure P** , if its performance at tasks in T , as measured by P , improves with experience E.”



Machine Learning

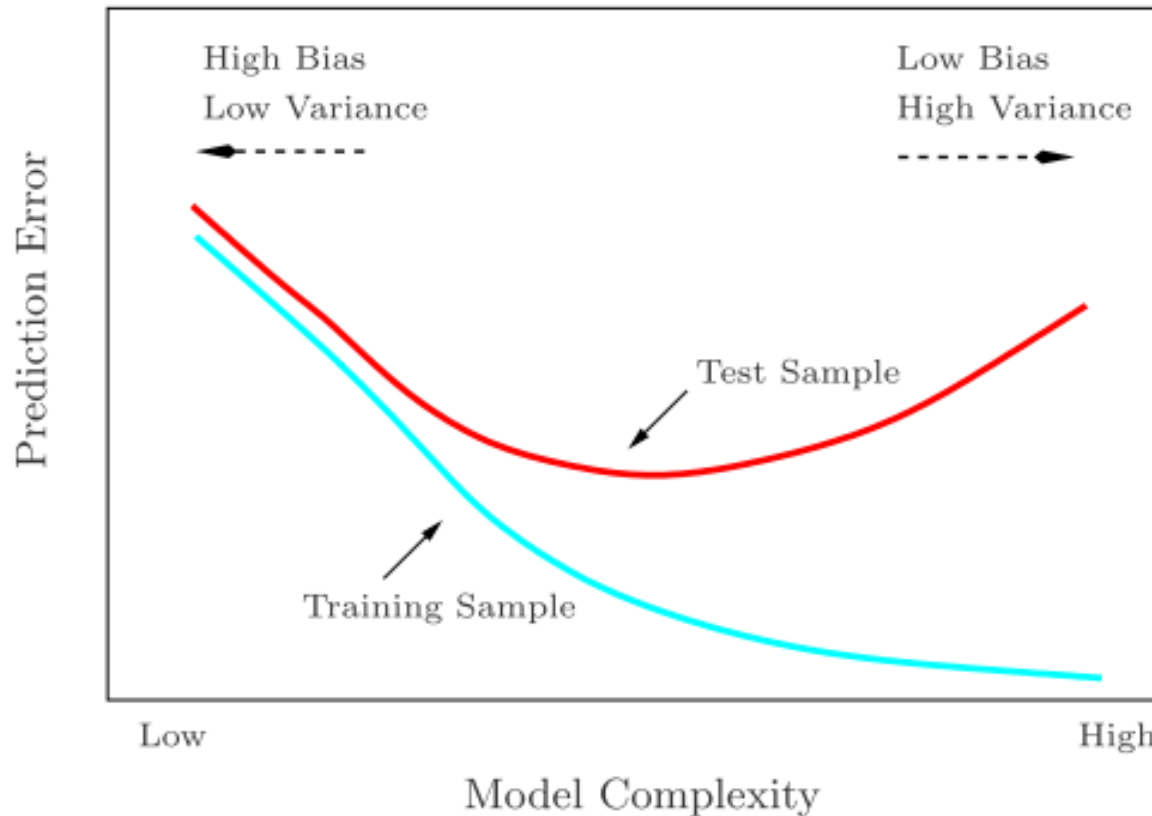
➤ Exemples de Tasks

- ☐ Classification
- ☐ Imputation of missing values
- ☐ Regression
- ☐ Denoising
- ☐ ... etc

➤ The Performance Measure, P

- ☐ Specific to the task
- ☐ Usually we are interested in how well the machine learning algorithm performs **on data that it has not seen before**, since this determines how well it will work when deployed in the real world.

Machine Learning



(Extrait de Hastie, Tibshirani, Friedman, *Statistical Learning* 2008)

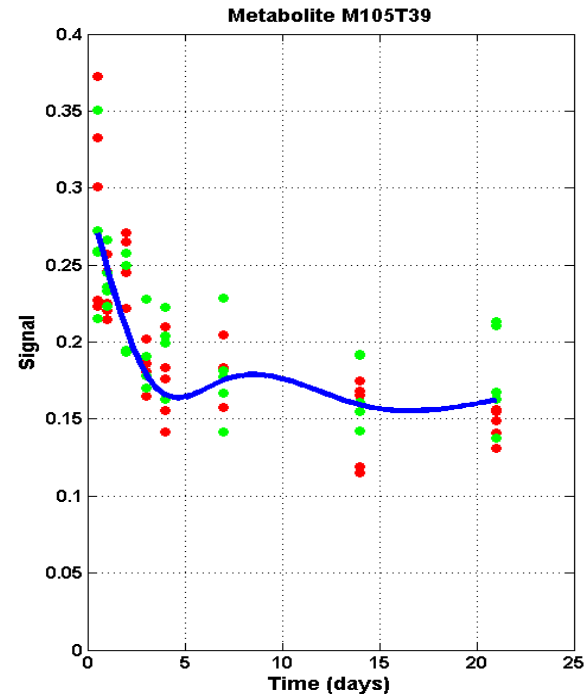
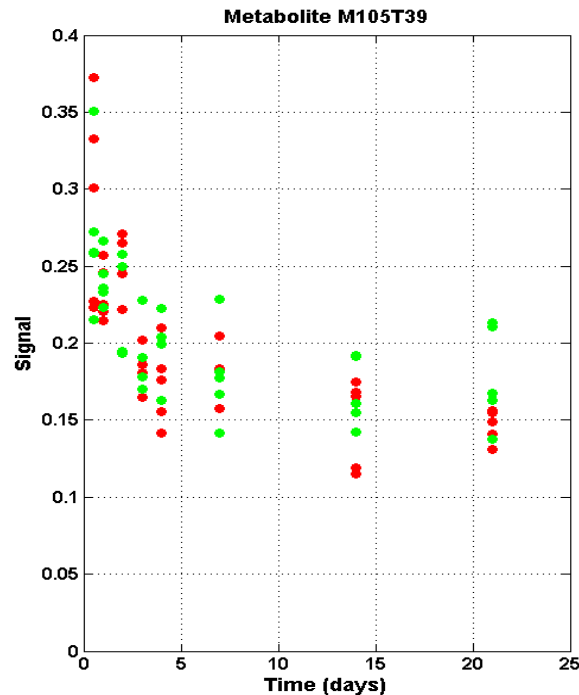
Rasoir d'Occam : principe de parcimonie

Pluralitas non est ponenda sine necessitate « Les multiples ne doivent pas être utilisés sans nécessité. »

Exemple 0

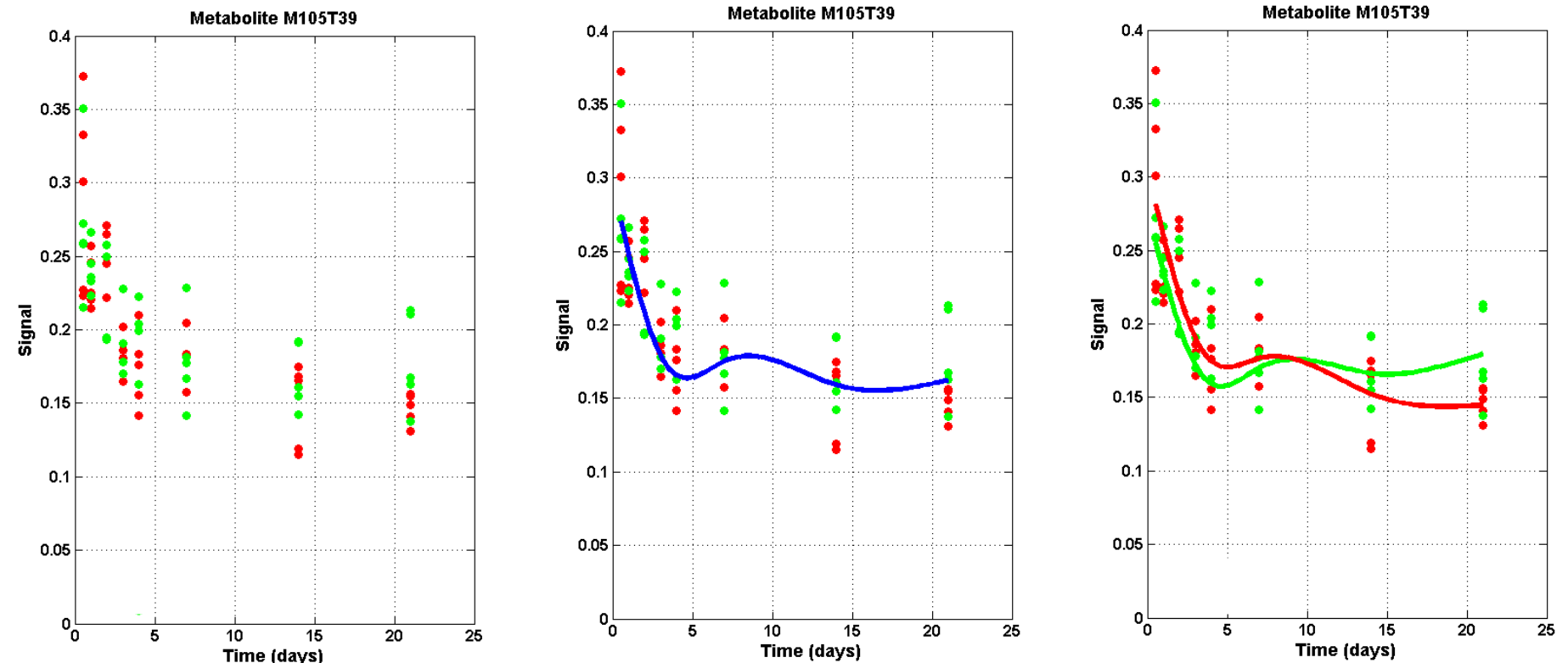
Des mesures discrètes vers les processus continu

Partie 1 : Analyse données Métabolomique de cellules Endotheliales



- Les données expérimentales peuvent être perçues comme une discrétisation de la cinétique d'un **processus continu**.
- Exemple de fonction spline pénalisée: $\|Y - X\beta\|^2 + \lambda\beta^T P\beta$
- Paramètre de lissage λ contrôle la parsimonie du modèle

Partie 1 : Analyse données Métabolomique de cellules Endotheliales



➤ Question fondamentale:

Dynamique unique sur le pool vs dynamique différenciée Irradiés vs contrôles ?

Partie 1 : Analyse données Métabolomique de cellules Endotheliales

- Question fondamentale:

Dynamique unique sur le pool vs dynamique différenciée Irradiés vs contrôles ?

- Il est possible de **complexifier le modèle**

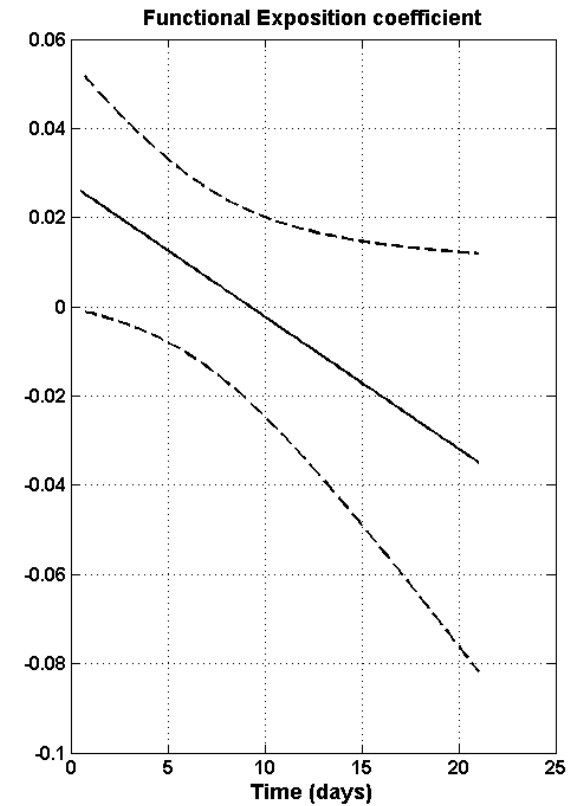
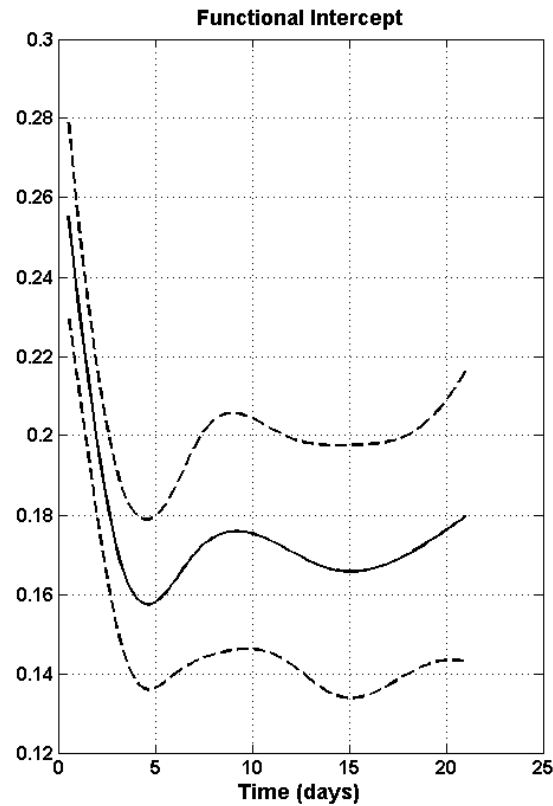
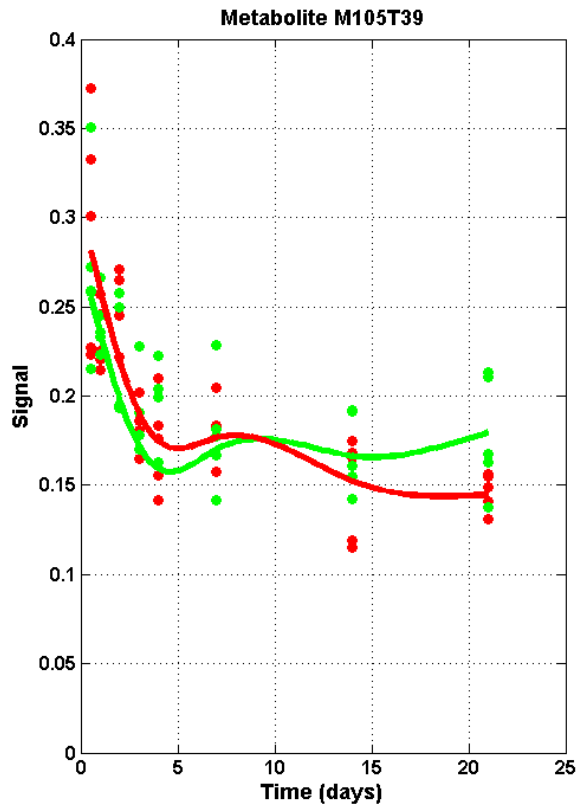
$$y_{ij} = \text{Intercept}(t_j) + \text{Coeff}(t_j) \times \text{Exposition}_i$$

Avec **Exposition = 0** pour les contrôles et **Exposition = 1** pour les exposés.

- Ce modèle **est-il « meilleur »** que le modèle plus simple poolé qui ne **dépend pas du statut exposé/contrôle**:

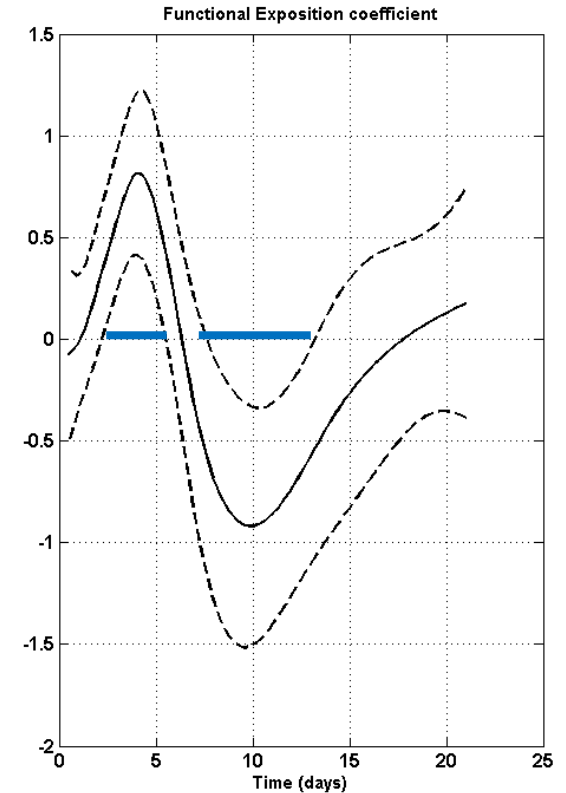
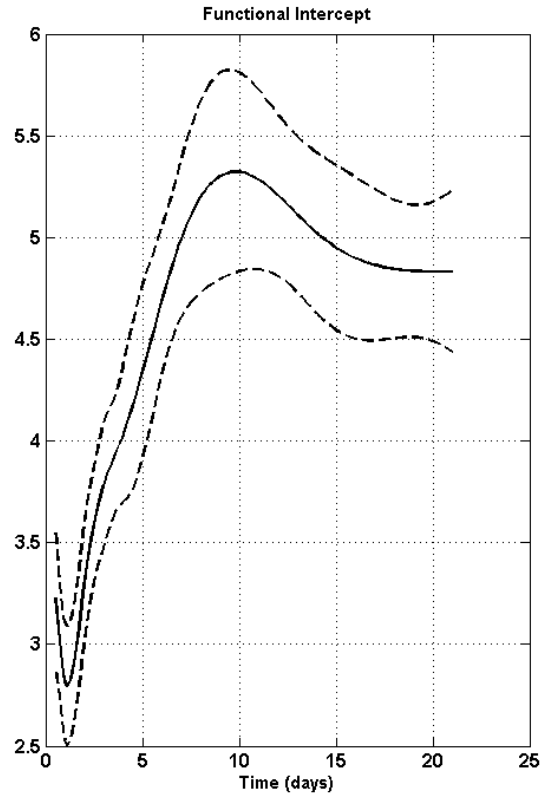
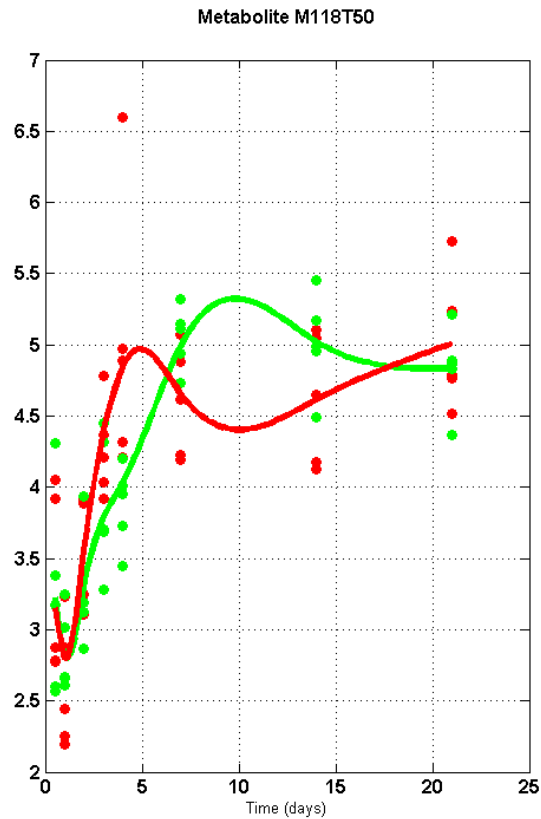
$$y_{ij} = \text{fonction_Pool}(t_j)$$

Partie 1 : Analyse données Métabolomique de cellules Endotheliales



$$\text{Fold Change} = \frac{\text{Signal exposé}}{\text{Signal Contrôle}} \Rightarrow \log_2(\text{FC}) = \log_2(\text{Signal Exposé}) - \log_2(\text{Signal Contrôle})$$

Partie 1 : Analyse données Métabolomique de cellules Endotheliales

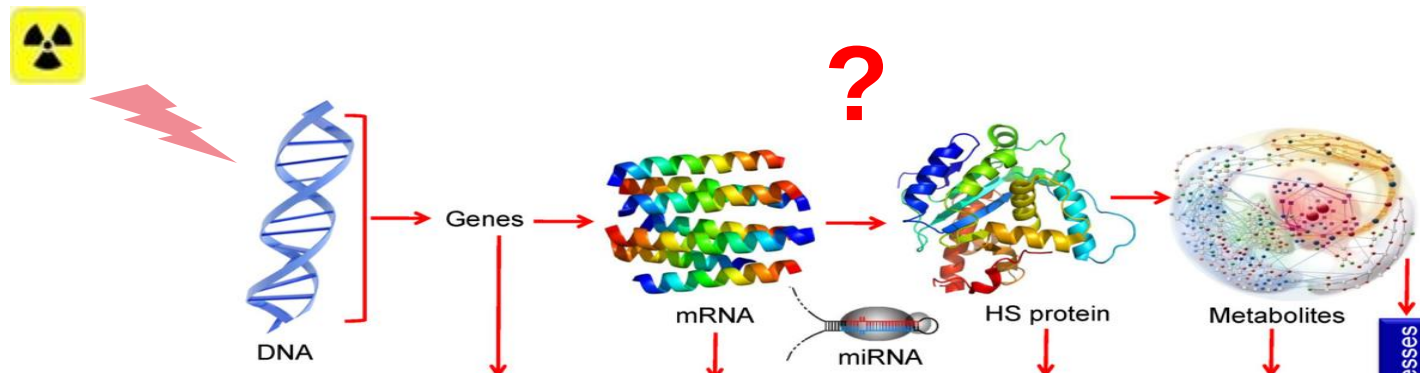


Exemple 1

**Intégration de réponses radio-induites à l'échelle
Transcriptomiques et protéomiques.**

Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques.

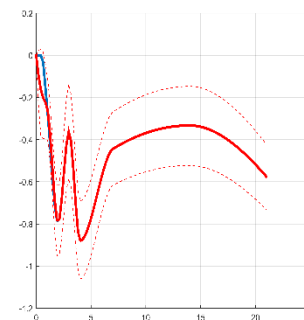
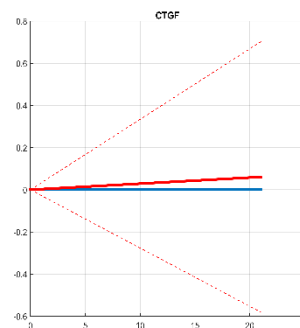
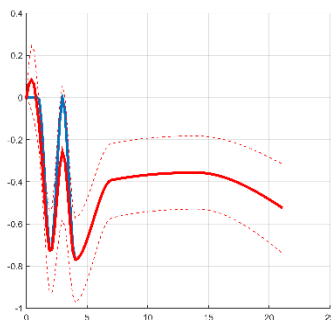
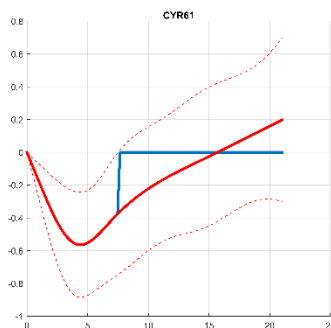
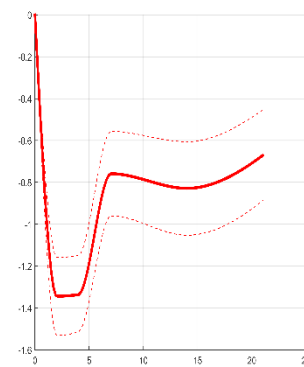
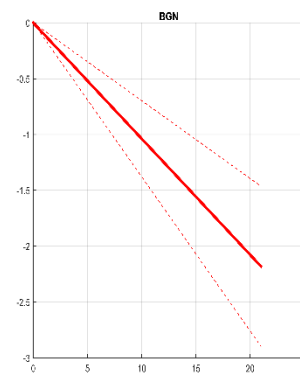
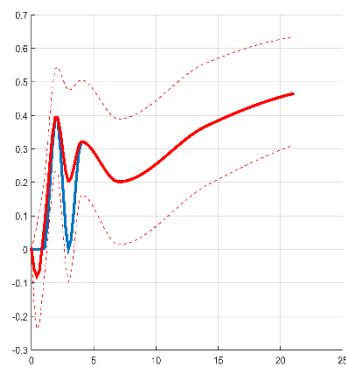
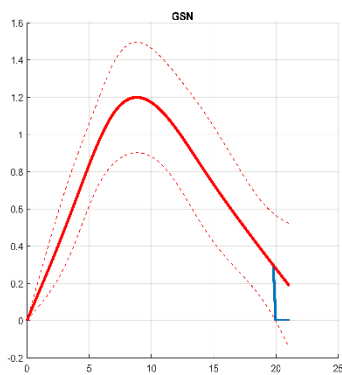
- To study **radiation effects**, modern transcriptomics and proteomics enable us to survey RNAs and protein expressions **at large scales**.
- These data still often generated and analyzed **separately** or under the assumption of a **matched correspondance** between mRNA abundance and alterations in protein expression which is **often not the case** [Rogers et al. *Bioinformatics*, 2008].
- In this work, we conducted **a simultaneous analysis** of radio-induced fold-changes in order to investigate how transcriptome and proteome expressions are linked and coordinated.



Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques

❑ Structure des données

- 7 pas de temps J1, J2 J3 J4 J7 J14 et J 21.
- 2 doses 0 et 20 Gy
- 4 réplicats par condition
- 86 Gènes avec une double cinétique transcriptomique (qPCR) et protéomique (spectroscopie de masse).



Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques: Que faire face à la malédiction dimensionnelle ?

- 86 couples d'observations dans un espace fonctionnelle de **dimension infinie**: Que faire face à la **malédiction dimensionnelle** ?

- Dimension reduction: Bivariate functional principal component analysis(FPCA)

It consists on a **Karhunen-Loève decomposition** [3] of the temporal bivariate functions

$f_i = (f_i^{Transcriptom}, f_i^{Proteom})$ into a **new feature coordinate system**:

$$f_i(t) = \bar{f}(t) + \sum_{k \geq 1} \theta_{ik} \times g_k(t) \quad (1)$$

Where $\bar{f}(t) = \frac{1}{n} \sum_{i=1}^n f_i(t)$ is the **common mean**

(θ_{ik}) are the uncorrelated (**scores**) with mean 0 and variances λ_k (**eigenvalues**)

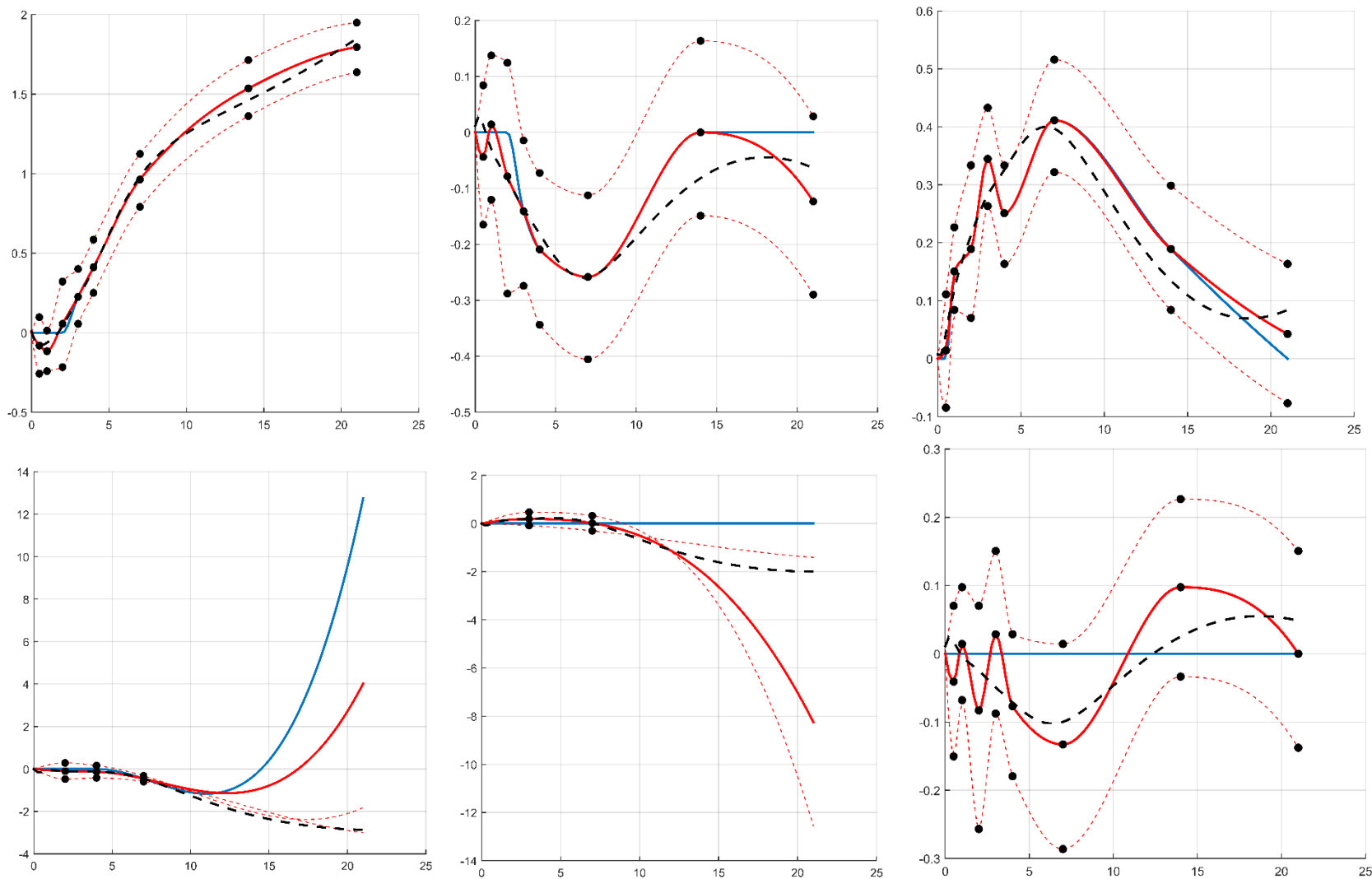
$g_k = (g_k^{Transcriptom}, g_k^{Proteom})$ are the **eigenfunctions** which describe the **main variation modes** within the data.

The truncature in (1) to the most important features provides **dimension reduction**.

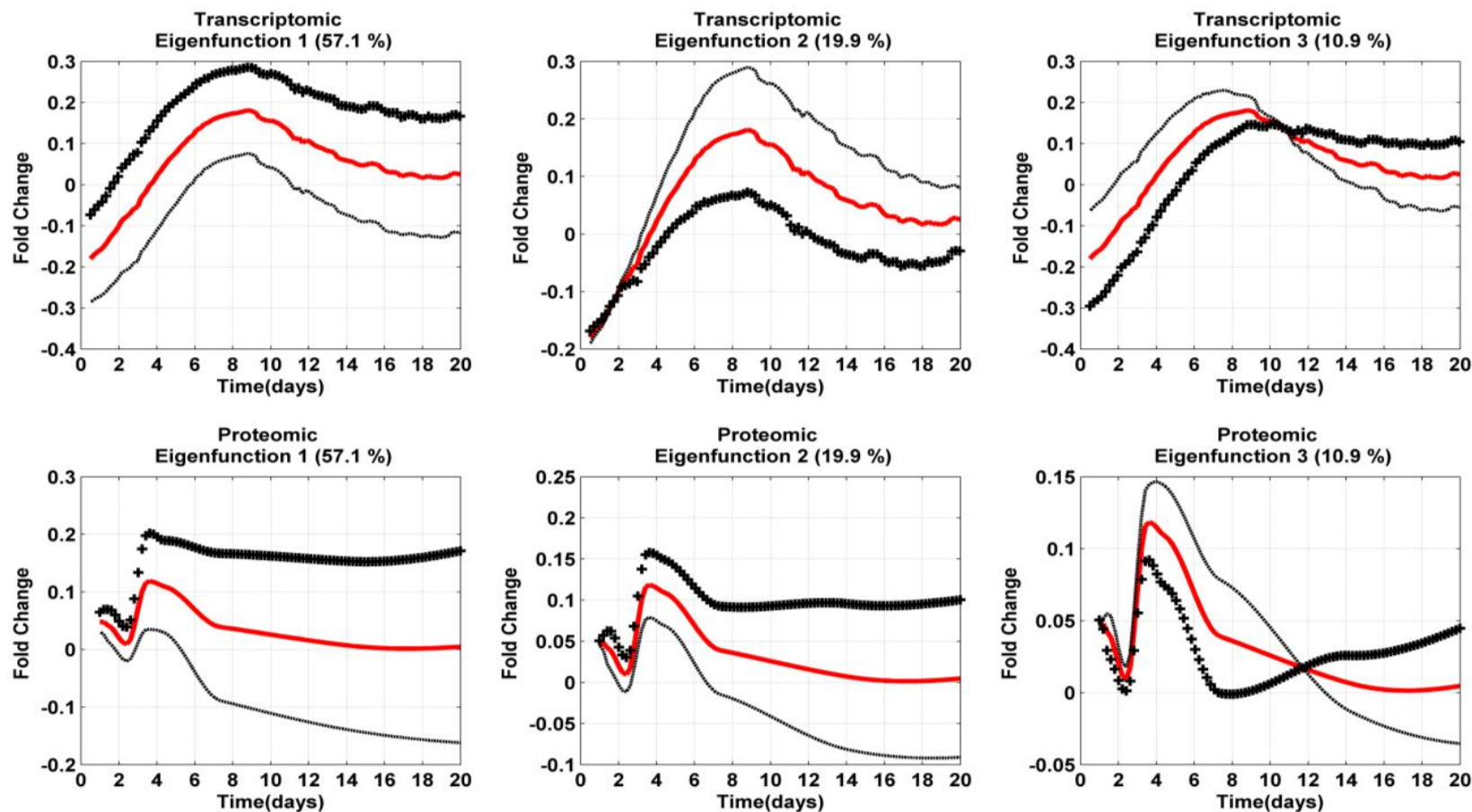
Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques: au départ 2500 Données protéomiques éparses !

	1	2	3	4	5	6	7	8	9
1	'A0A024R3N8'	NaN	NaN	NaN	NaN	0.9070	NaN	NaN	NaN
2	'A0A024R3N8'	NaN	NaN	NaN	NaN	0.7470	NaN	NaN	NaN
3	'A0A024R3N8'	NaN	NaN	NaN	NaN	1.1000	NaN	NaN	NaN
4	'A0A024R462'	NaN	0.8930	0.8930	0.7780	0.8270	0.6220	0.6600	0.5750
5	'A0A024R462'	NaN	0.7700	0.7120	0.6500	0.6930	0.5060	0.5360	0.4570
6	'A0A024R462'	NaN	1.0400	1.1100	0.9300	0.9860	0.7750	0.8160	0.7210
7	'A0A024R4N1'	NaN	NaN	NaN	1.0400	NaN	NaN	NaN	NaN
8	'A0A024R4N1'	NaN	NaN	NaN	0.8450	NaN	NaN	NaN	NaN
9	'A0A024R4N1'	NaN	NaN	NaN	1.2700	NaN	NaN	NaN	NaN
10	'A0A024RAB6'	0.9430	NaN	0.9590	0.9160	0.9410	0.8840	0.7650	0.6300
11	'A0A024RAB6'	0.8170	NaN	0.7600	0.7730	0.7770	0.7160	0.6220	0.5020
12	'A0A024RAB6'	1.0800	NaN	1.2100	1.0800	1.1400	1.0800	0.9350	0.7840
13	'A0A024RDE8'	1.0800	1.0200	NaN	1.0300	NaN	NaN	NaN	NaN
14	'A0A024RDE8'	0.9300	0.8900	NaN	0.8650	NaN	NaN	NaN	NaN
15	'A0A024RDE8'	1.2600	1.1700	NaN	1.2200	NaN	NaN	NaN	NaN
16	'A0A087X0G9'	NaN	NaN	NaN	NaN	1.0700	NaN	NaN	NaN
17	'A0A087X0G9'	NaN	NaN	NaN	NaN	0.8630	NaN	NaN	NaN
18	'A0A087X0G9'	NaN	NaN	NaN	NaN	1.3100	NaN	NaN	NaN
19	'A0A0B5J5S1'	NaN	NaN	NaN	NaN	NaN	NaN	0.3590	NaN
20	'A0A0B5J5S1'	NaN	NaN	NaN	NaN	NaN	NaN	0.2820	NaN
21	'A0A0B5J5S1'	NaN	NaN	NaN	NaN	NaN	NaN	0.4600	NaN
22	'A0A0F7G8J1'	NaN	NaN	NaN	NaN	1.0600	NaN	NaN	NaN
23	'A0A0F7G8J1'	NaN	NaN	NaN	NaN	0.8530	NaN	NaN	NaN
24	'A0A0F7G8J1'	NaN	NaN	NaN	NaN	1.3000	NaN	NaN	NaN

Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques: Réduction dimensionnelle et données éparées



Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques : la FPCA



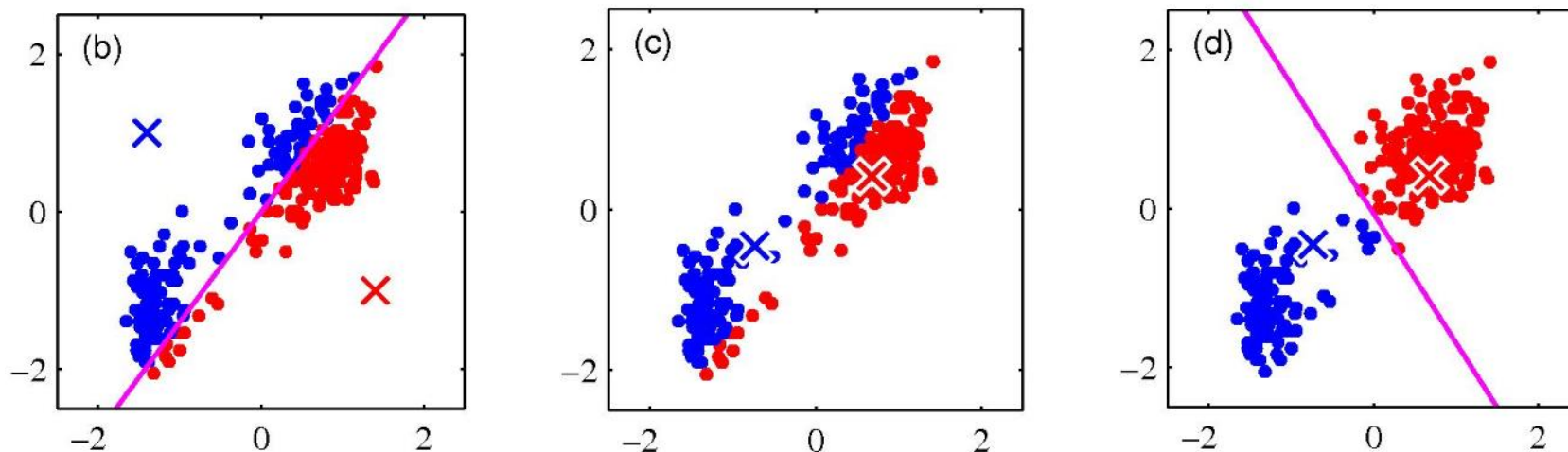
Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques : Clustering des couples de profils temporelles synchronisés [Sangalli et al, mean alignment for curve clustering 2010]

► K-means clustering for time-warped bivariate temporal profiles [4]

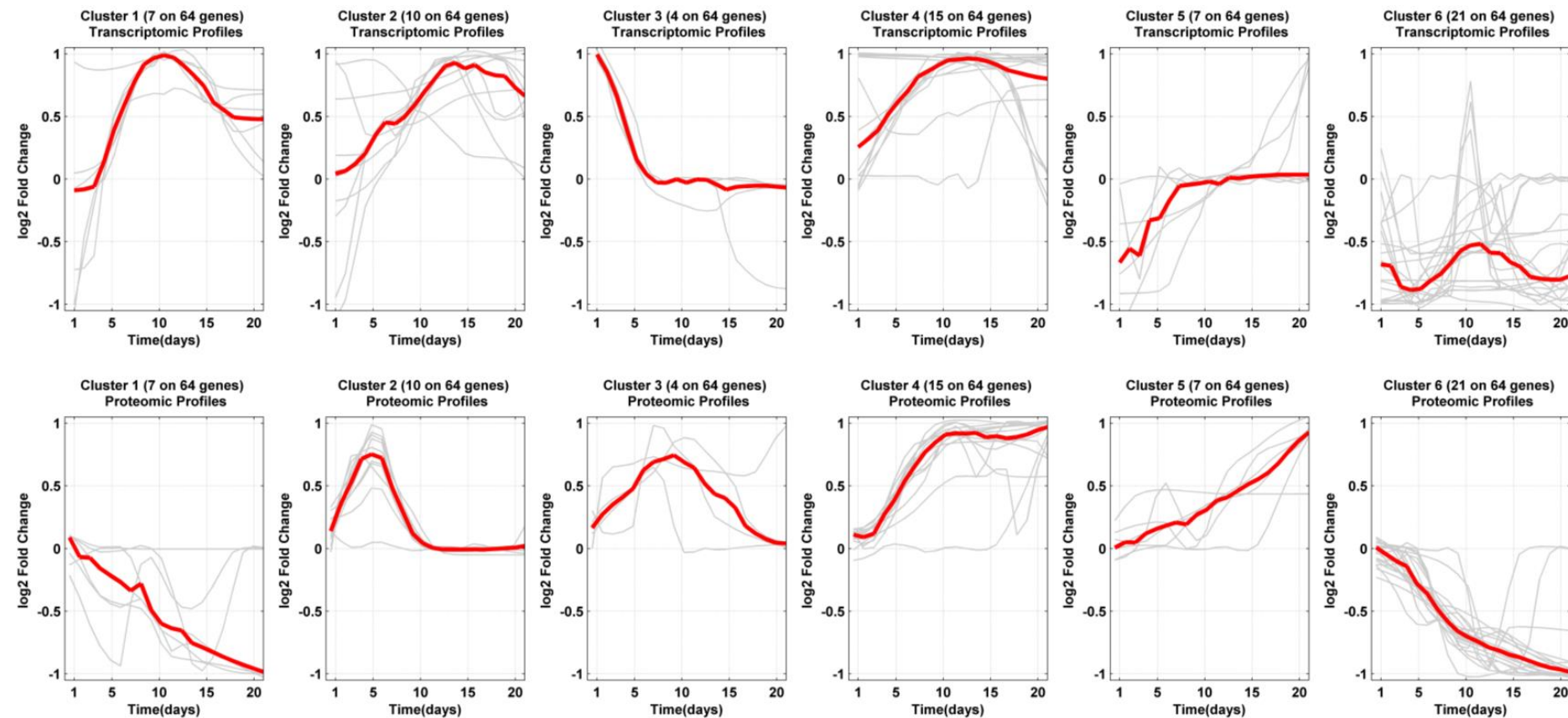
Since the temporal trajectories reflects variation both in **amplitudes** and temporal **phases**, it seems natural to perform the **clustering and the time-synchronization simultaneously**.

Given K template curves $(\varphi_1, \dots, \varphi_K)$, the **alignment step** of the algorithm consists on the time-warping $\left(\tilde{f}_{ik} = f_i \circ h_{ik}\right)_{k=1..K}$ of each (f_i) to the $(\varphi_1, \dots, \varphi_K)$ prior to the **assignment step**. At each iteration, the warping functions (h_{ik}) **are averaged** to $\bar{h}_k$ within the k^{th} cluster in order **to update** the aligned function by $\tilde{f}_i = \tilde{f}_i \circ \bar{h}_k^{-1}$. The templates $(\varphi_1, \dots, \varphi_K)$ can then be updated by **averaging** $\left(\tilde{f}_i\right)$ **within each cluster**.

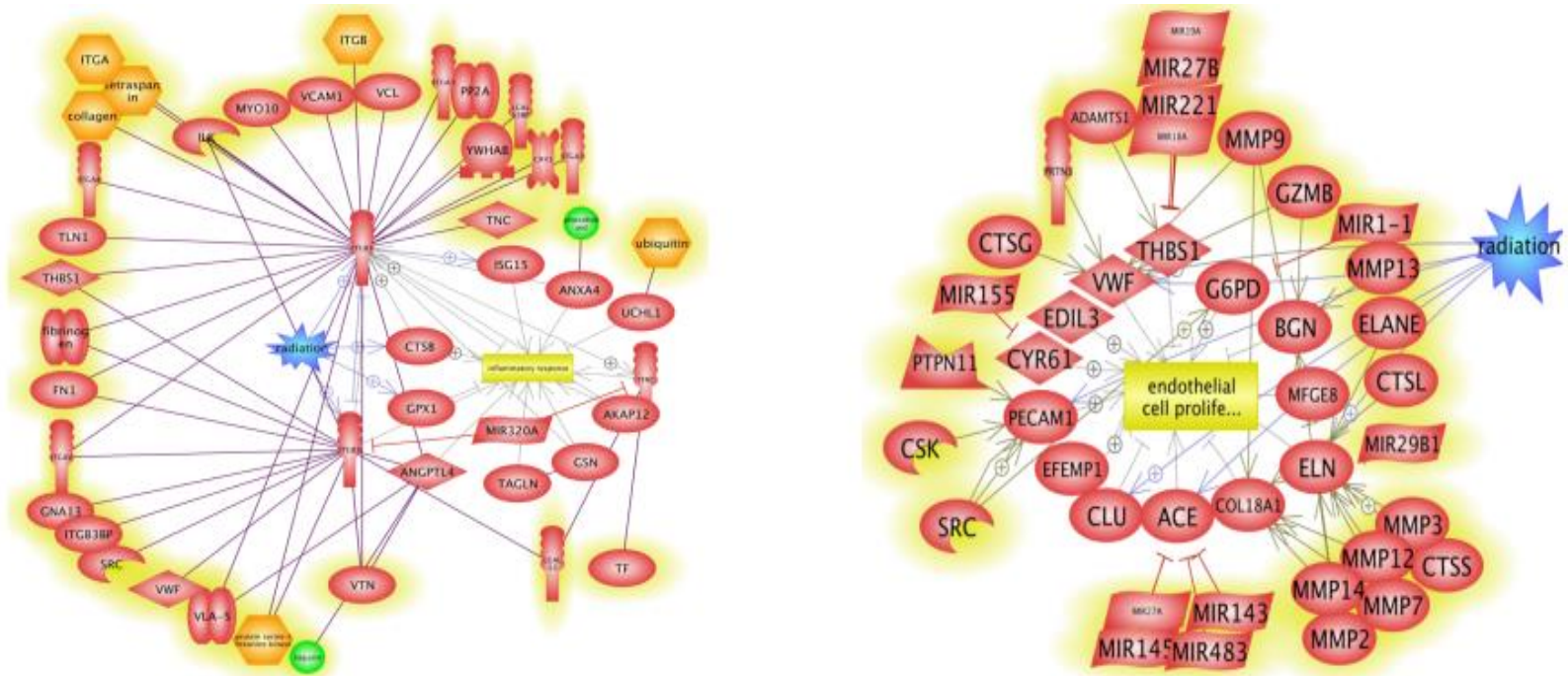
Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques : Principe de l'algorithme K-means



Exemple 1: Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques : Clustering des couples de profils temporels synchronisés [*Sangalli et al, mean alignment for curve clustering 2010*]



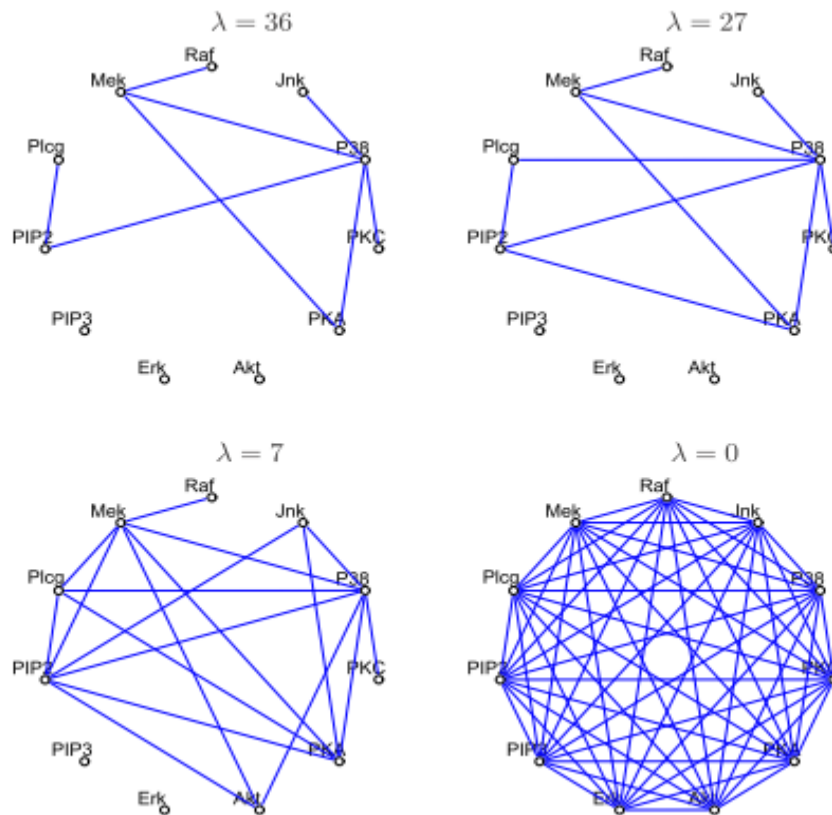
Exemple 1: Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques : Clustering des couples de profils temporelles synchronisés [Sangalli et al, mean alignment for curve clustering 2010]



- An enrichment analysis shows that a **high correlation** of mRNA and protein radio-induced expression occurs in mechanisms of **inflammatory response** and inhibition of **cell proliferation**.

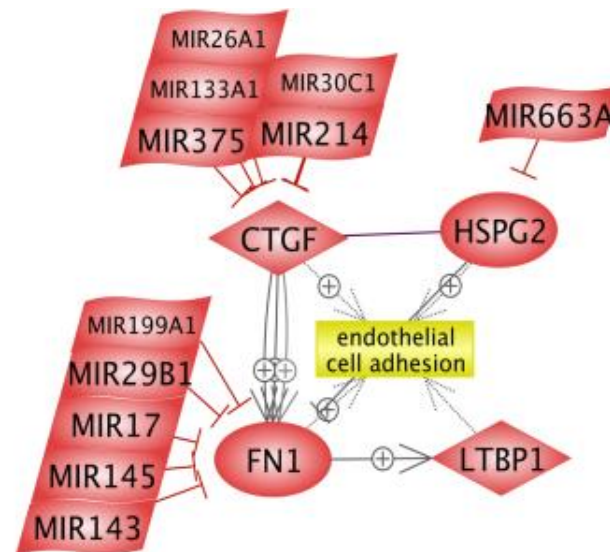
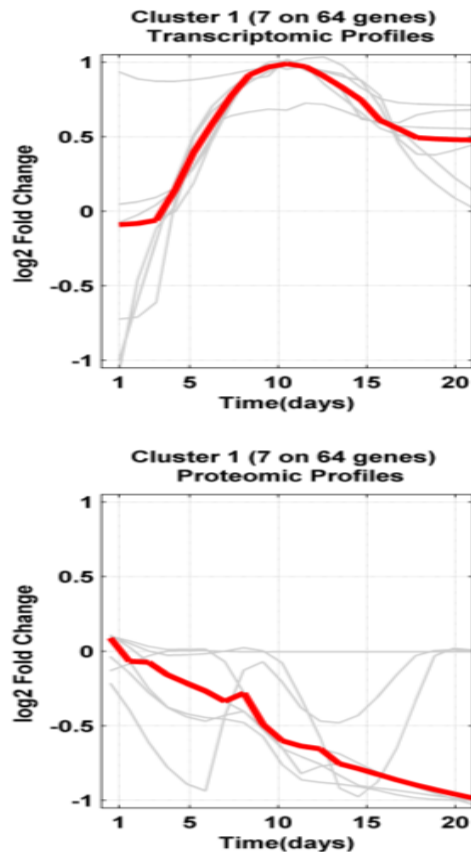
Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques : Inférer la structure du graph

- Objectif: Trouver la **structure** du graph.
- Méthode: regression **alternée** d'une entitée **par toutes les autres**.
- Selection de variable **drastique** pour sélectionner les voisins les plus connectées (régression lasso)



Extrait de Hastie, Tibshirani, Friedman. Statistical learning 2008

Exemple 1: Intégration de réponses radio-induites à l'échelle Transcriptomiques et protéomiques : Clustering des couples de profils temporelles synchronisés [Sangalli et al, mean alignment for curve clustering 2010]



Exemple 2

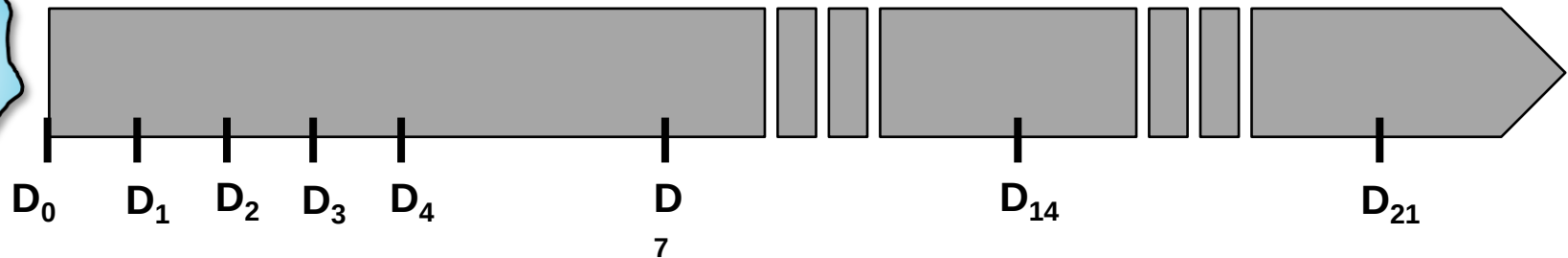
Analyse de données d'expression de gènes en dynamique temps et dose

HUVEC's dynamic molecular profile

HUVEC

Doses : 0, 1, 2, 3, 4, 5, 6, 10 et 20 Gy

n = 4 réplicats



Total RNA



Gene expression assay

44 gènes impliqués dans la senescence

+ 4 gènes ménage

HUVEC's dynamic molecular profile

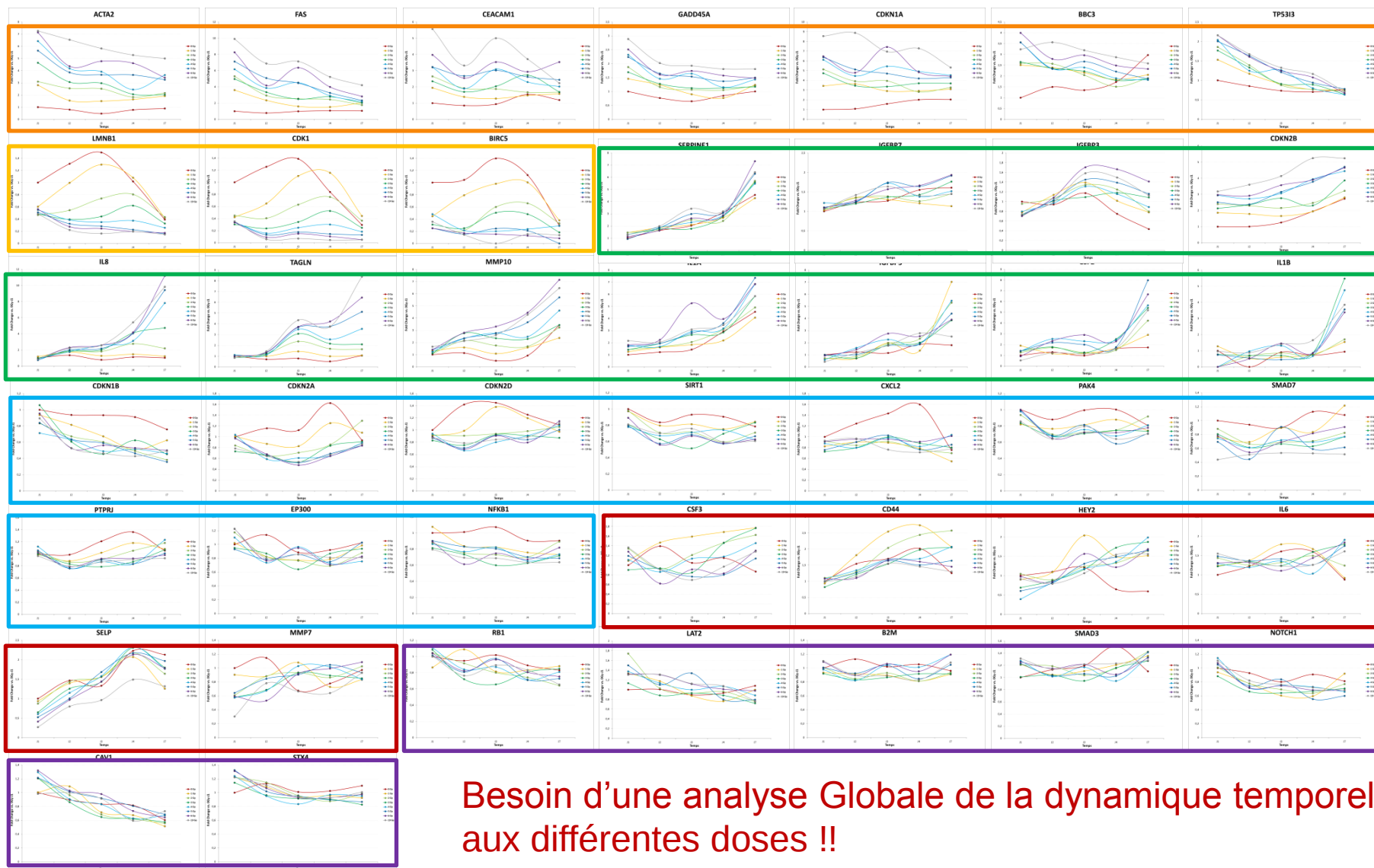
- ❑ *Structure des données*

- 7 pas de temps J1,J2 J3 J4 J7 J14 et J 21.
- 9 doses 0 ,1,2,3,4,5,6,10,et 20 Gy
- 4 réplicats par condition

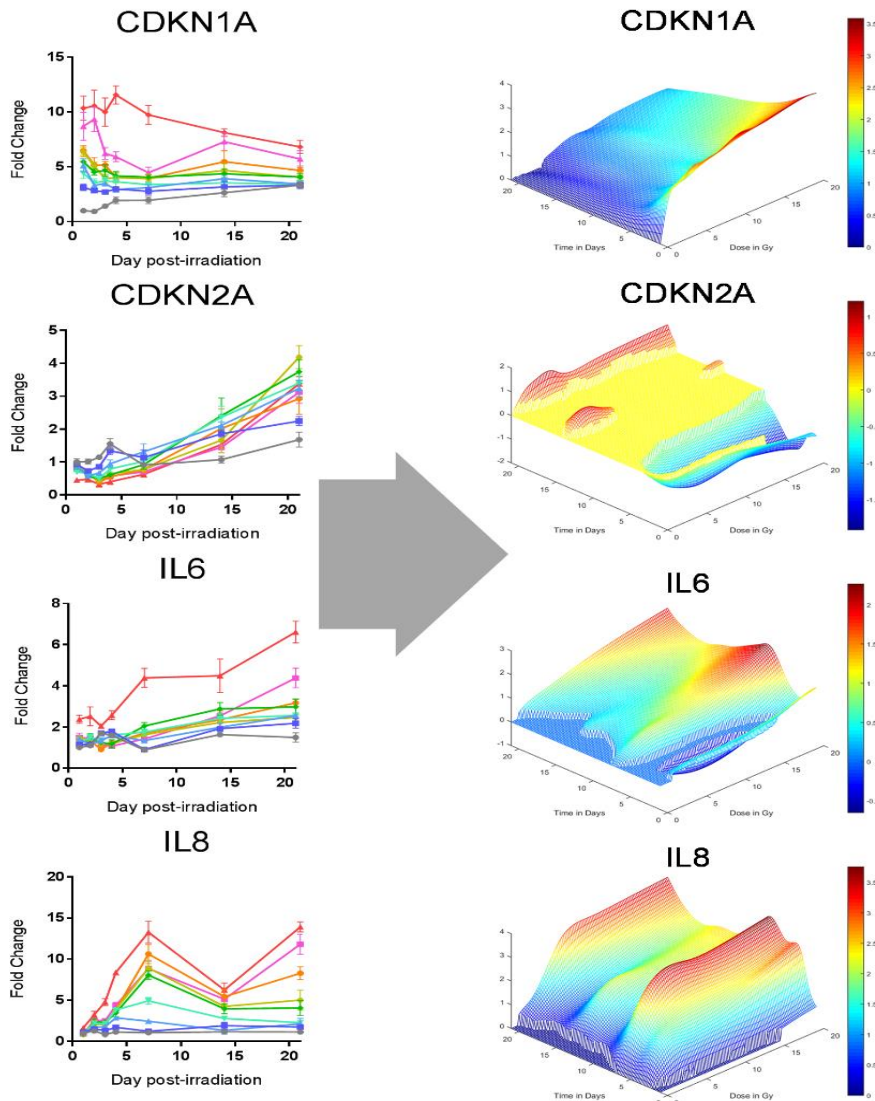
- ❑ *Au total 252 différentes conditions réplicats compris pour chaque gène.*

- ❑ 44 Gènes impliqués dans la sénescence.
- ❑ 4 Gènes de ménage.

Results : HUVEC's dynamic molecular profile



Besoin d'une analyse Globale de la dynamique temporelle aux différentes doses !!



➤ Representation en processus bivarié en temps et en dose:

$$f(\delta_i, t_j) = \sum_{k=1}^K \sum_{l=1}^L \beta_{lk} \times b_k^{\text{Dose}}(\delta_i) \times b_l^{\text{Time}}(t_j)$$

Where $\beta = (\beta_{kl})_{1 \leq k \leq K, 1 \leq l \leq L}$ represents the coefficient vector of the expansion of f in the tensor B-spline basis.

We minimize the penalized sum of squared errors criterion :

$$\|Y - X\beta\|^2 + \beta^T P \beta$$

Where:

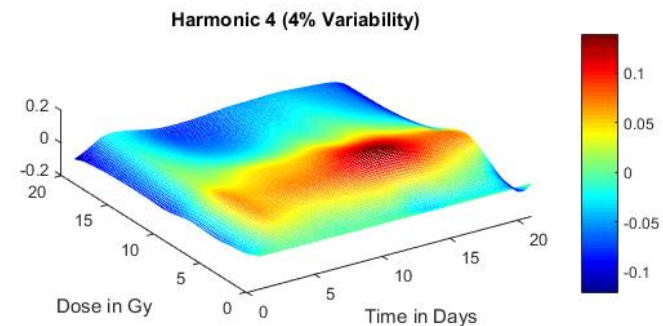
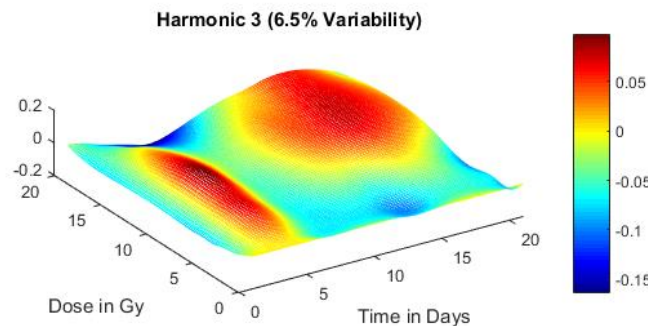
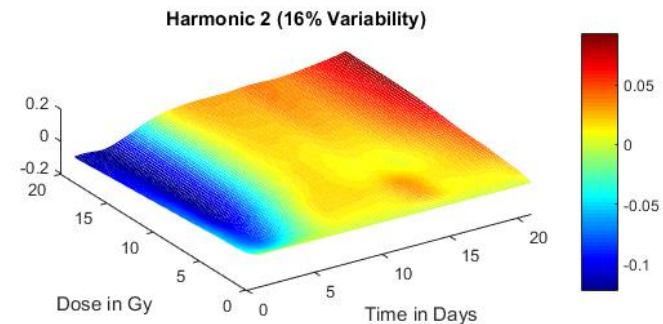
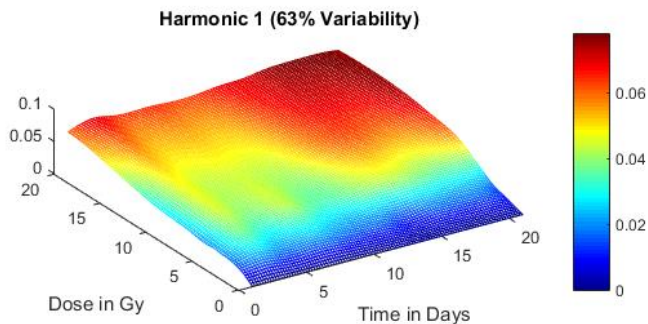
- $Y = (-\Delta C_{Tij})_{1 \leq i \leq n, 1 \leq j \leq m}$
- $X = (B_k(\delta_i, t_j))_{1 \leq i \leq n, 1 \leq j \leq m, 1 \leq k \leq K}$
- P the tensor product of marginal second derivative penalties :

$$P = \lambda_1 \times P^{\text{Dose}} \otimes I_L + \lambda_2 \times I_K \otimes P^{\text{Time}}$$

Using the FPCA eigenfunctions, it's now possible to represent each gene's surface profile in term of the Karhunen-Loève decomposition (Karhunen 1947)

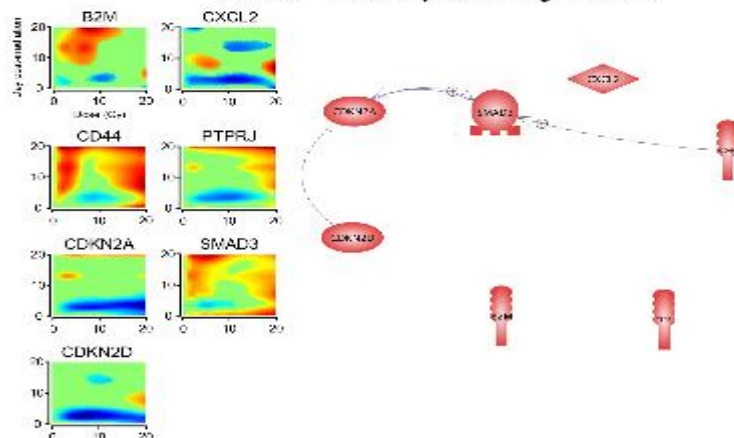
$$f_i(\delta, t) = \bar{f}(\delta, t) + \sum_{j \geq 1} \xi_{ij} \times \phi_j(\delta, t) \approx \tilde{f}_i(\delta, t) = \bar{f}(\delta, t) + \sum_{j=1}^J \xi_{ij} \times \phi_j(\delta, t)$$

where $\bar{f}(\delta, t) = \frac{1}{n} \sum_{i=1}^n f_i(\delta, t)$ is the common mean and (ξ_{ik}) are uncorrelated random effect variables (scores) with mean 0 and variances λ_k (eigenvalues)

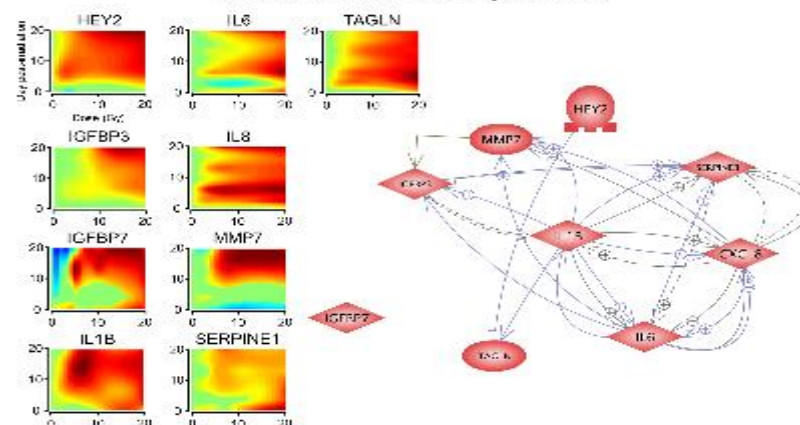


HUVEC's dynamic molecular profile

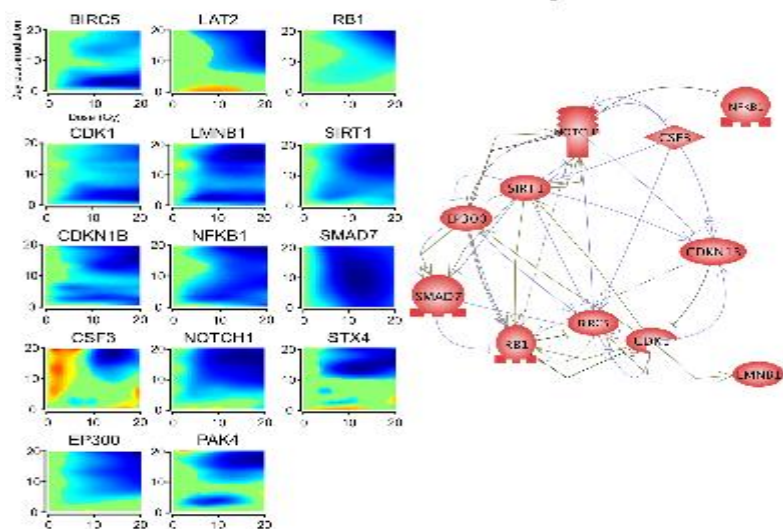
Cluster 1 « Early downregulation »



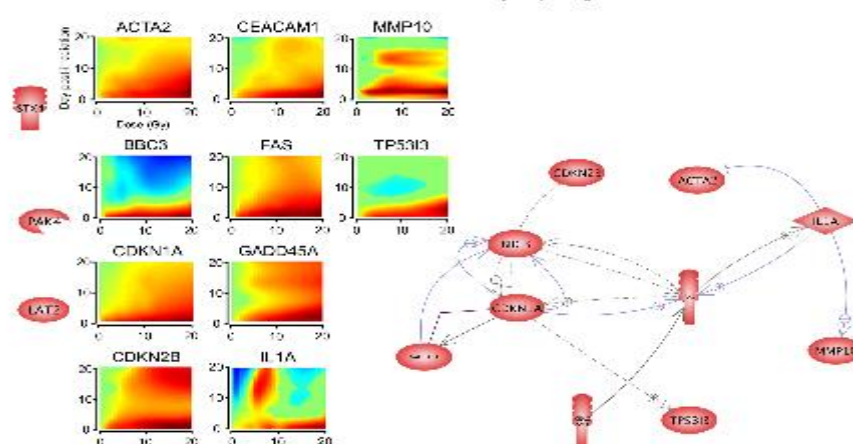
Cluster 2 « Late upregulation »



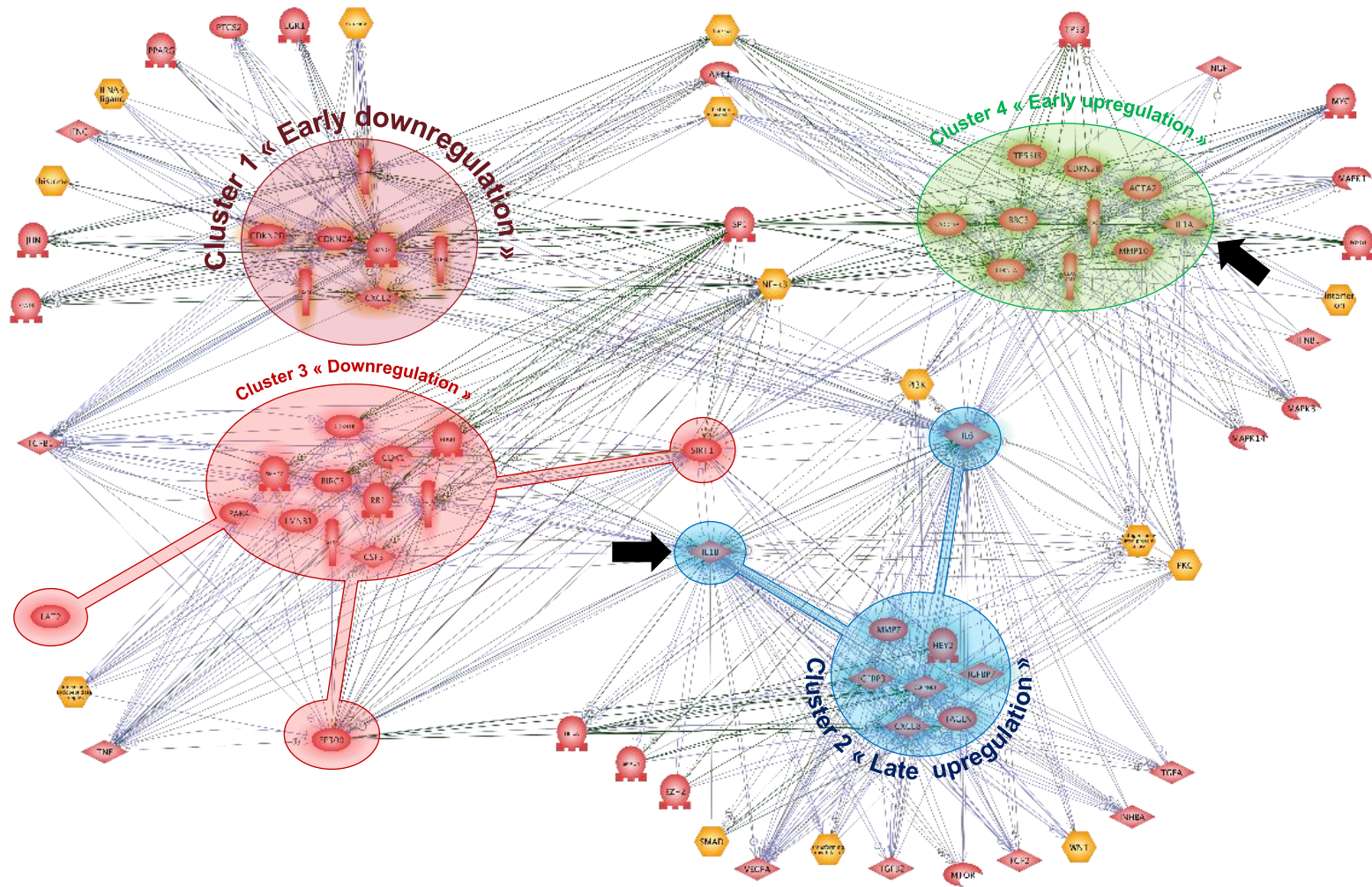
Cluster 3 « Downregulation »



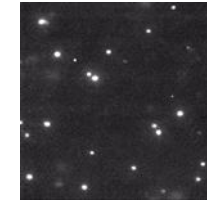
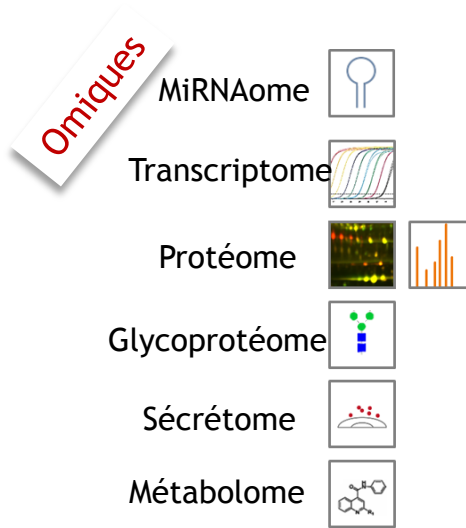
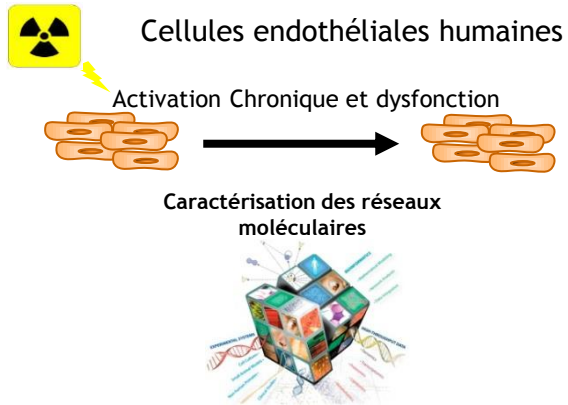
Cluster 4 « Early upregulation »



HUVEC's dynamic molecular profile



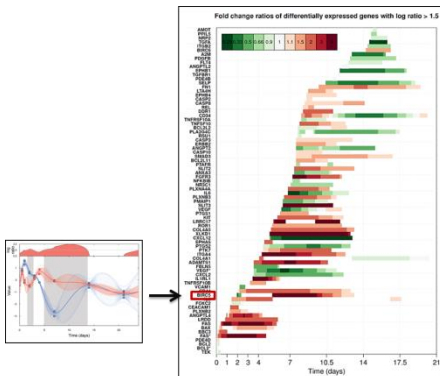
Conclusion



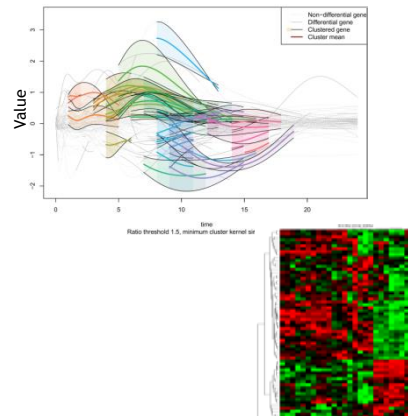
Fonction cellulaire

Réduire la dimension des données pour les « faire parler »

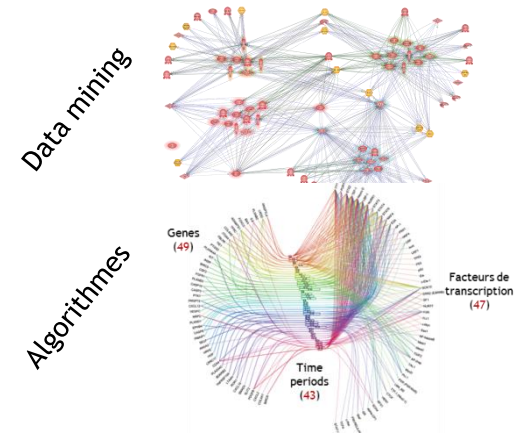
Modèles **mathématiques** appliqués aux dynamiques temporelles



Outils **statistiques** de clustérsation



Inférence des réseaux moléculaire par la **bio-informatique**



Finalité : Trouver des nœuds moléculaires impliqués dans la dysfonction pour validation potentielle *in vivo*
 Trouver des Variables biologiques pour mieux évaluer le risque de dysfonction cellulaire

Merci à



Dr Olivier Guipaud (LRMED)

Dr Jean-Philippe Chan
Dr David Vaudry

Dr Jean-Charles Martin

Frédéric Soysouvanh (thèse
Senescence LRMED)

Dr Fabien Milliat (chef du
Laboratoire LRMED)

Dr Marc Benderitter (chef du
SERAMED)



Électricité de France

Merci pour votre attention

Question ?