Comparaison de 5 méthodes de machine learning non supervisées appliquées à des données de métabolomique chez des patientes atteintes d'un cancer du sein

Jocelyn Gal[1*], David Chardin[2,3], Caroline Bailleux[2], Thierry Pourcher[4], Jean-Marc Ferrero[2], Emmanuel Barranger[5], Olivier Humbert[2,3], Emmanuel Chamorey[1]

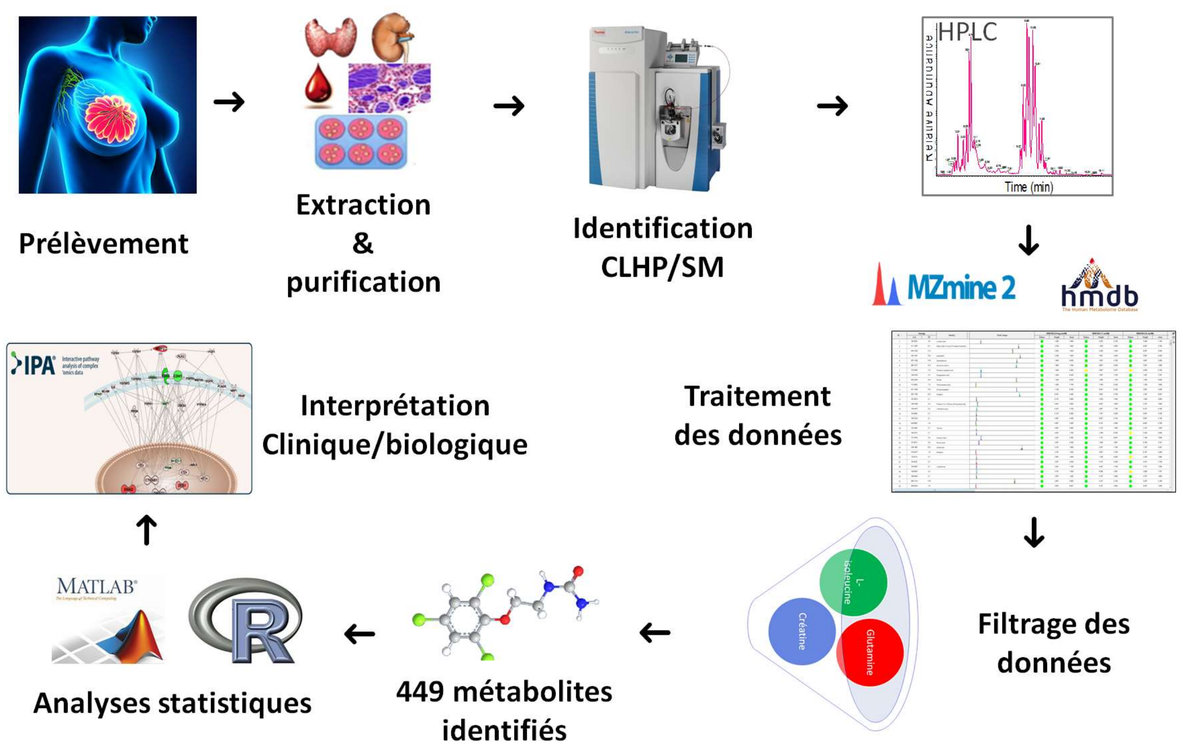
1-Département d'Epidémiologie et Biostatiques, Centre Antoine Lacassagne, Nice; 2-Département d'oncologie médical, Centre Antoine Lacassagne, Nice; 3- Département de médecine nucléaire, Centre Antoine-Lacassagne, Nice; 4- Université Côte-d'Azur, Laboratoire TIRO, Faculté de médecine, Nice; 5- Département de chirurgie, Centre Antoine-Lacassagne, Nice

Introduction

La transcriptomique a permis la classification des cancers du sein, elle est très largement utilisée en pratique clinique courante. Cependant, il persiste une hétérogénéité des sous-types de cancer du sein qui souligne la nécessité d'affiner cette classification. La métabolomique est un domaine en pleine expansion dédié à l'étude du métabolisme. L'objectif de cette étude était d'identifier des signatures métabolomiques obtenues à l'aide de 5 méthodes de machine learning (ML) non supervisées.

Méthodes

Figure 1: Processus d'une analyse en métabolomique non ciblée



- **52 patientes** atteintes d'un cancer du sein et traitées de façon **adjuvante** entre 2013 et 2016.
- **449 métabolites** extraits par CLHP/SM (**Figure 1**)
- Méthodes de ML non supervisées comparées: **PCA K-means; Sparcl; Spectral clustering; SIMLR et K-sparse**.
- Méthodes pour définir le nombre optimal de clusters: **GAP statistic**.
- Méthode d'évaluation de la performance: **indice de silhouette**.

Résultats

- **Caractéristiques des patientes** : âge moyen 63,2 ans; T1/T2: 45 patientes (86,5%); SBR III: 24 patientes (47%); N+: 24 patientes (46%); HER2+: 12 patientes (23%); Triple-négatif: 15 patientes (29%); Luminal A/B: 25 patientes (48%); RH+: 27 patientes (52%).
- **Suivi médian**: 48.5 mois (IC à 95% [43-54,5]).
- **Survie globale à 3 ans :** 90% (IC 95% [82-99])
- **Survie sans récidive à 3 ans:** 90% (IC95% [82-99])
- **Nombre optimal de clusters**: K=3

Tableau 1: Performance des 5 méthodes.

Méthodes	Silhouette
K-Sparse	**0,91**
SIMLR	**0,85**
Sparse Kmeans	0,72
Spectral clustering	0,64
PCA k-means	0,26

- Les méthodes: SIMLR et K-sparse ont été plus performantes que les 3 autres.

Tableau 2: Comparaison des caractéristiques clinique entre les 3 clusters obtenus avec K-Sparse et SIMLR

	SIMLR				K-Sparse			
	Cluster 1 N=17 (%)	Cluster 2 N=12 (%)	Cluster 3 N=23 (%)	p	Cluster 1 N=19 (%)	Cluster 2 N=12 (%)	Cluster 3 N=21 (%)	p
T				**0,045**				**0,018**
T1/T2	10 (58,8)	6 (50)	5 (21,7)		12 (63,2)	5 (41,7)	4 (19)	
T3	7 (41,2)	6 (50)	18 (78,3)		7 (36,8)	7 (58,3)	17 (81)	
SBR				**0,007**				**0,025**
I/II	11 (64,7)	9 (75)	7 (31,8)		12 (63,2)	9 (75)	6 (30)	
III	6 (35,3)	3 (25)	15 (68,2)		7 (36,8)	3 (25)	14 (70)	
Ki67	38 (31)	32.8 (22,7)	59.7 (25,9)	**0,009**	41.1 (30,6)	33 (22,6)	58.8 (27,2)	**0,027**
Phenotype				**0,018**				**0,012**
HER2	1 (5,9)	4 (33,3)	7 (30,4)		1 (5,3)	4 (33,3)	7 (33,3)	
Luminal	12 (70,6)	7 (58,3)	6 (26,1)		13 (68,4)	7 (58,3)	5 (23,8)	
TN	4 (23,5)	1 (8,3)	10 (43,5)		5 (26,3)	1 (8,3)	9 (42,9)	

Les méthodes K-sparse et SIMLR ont généré 3 clusters avec des différences cliniques statistiquement significatives, non superposables avec les sous-types histologiques connus.

Conclusion / Perspective

Cette étude a permis, à l'aide de méthodes de ML non supervisées, d'identifier 3 groupes de patientes présentant des caractéristiques clinico-biologiques distinctes, à partir de données de métabolomiques. La validation de nos résultats sur une cohorte externe est nécessaire. La métabolomique semble être un outil pertinent et prometteur pour la classification des cancers du sein.