Développement d'une méthodologie robuste de sélection de gènes dans le cadre d'une activation pharmacologique de la voie PPAR

Aurélie Cotillard

ÉCOLE CENTRALE DES ARTS ET MANUFACTURES « ÉCOLE CENTRALE PARIS »

THÈSE
présentée par

Aurélie COTILLARD

pour l'obtention du

GRADE DE DOCTEUR

Spécialité : Mathématiques appliquées

Laboratoire d'accueil : MAS

SUJET : Développement d'une méthodologie robuste de sélection de gènes dans le cadre d'une activation pharmacologique de la voie PPAR

soutenue le : 03/12/09

devant un jury composé de :

M. Jean-Philippe Vert	**Président**
M. Pascal Barbry	**Rapporteur**
M. Avner Bar-Hen	**Rapporteur**
M. Jean-Pierre Galizzi	**Examinateur**
Mme Françoise Xavier	**Examinateur**
M. Christian Saguez	**Directeur de thèse**
M. Brian Lockhart	**Invité**

2009ECAP0040

Résumés

Résumé en français :

De part leur dimension élevée, les données de puces à ADN nécessitent l'application de méthodes statistiques pour en extraire une information pertinente. Dans le cadre de l'étude des différences entre deux agonistes de PPAR (Peroxisome Proliferator-Activated Receptor), nous avons sélectionné trois méthodes de sélection de variables : T-test, Nearest Shrunken Centroids (NSC) et Support Vector Machine – Recursive Feature Elimination. Ces méthodes ont été testées sur des données simulées et sur les données réelles de l'étude PPAR. En parallèle, une nouvelle méthodologie, MetRob, a été développée afin d'améliorer la robustesse ce ces méthodes vis à vis de la variabilité technique des puces à ADN, ainsi que leur reproductibilité. Cette nouvelle méthodologie permet principalement d'améliorer la valeur prédictive positive, c'est-à-dire la confiance accordée aux résultats. La méthode NSC s'est révélée la plus robuste et ce sont donc les résultats de cette méthode, associée à MetRob, qui ont été étudiés d'un point de vue biologique.

Mots clés : Puces à ADN, Sélection de variables, Traitement de données, PPAR, Diabète de type 2

Résumé en anglais :

The microarray technology provides high dimensional data that need to be statistically treated for extracting relevant information. Within the context of the study of the differences between two PPAR (Peroxisome Proliferator-Activated Receptor) agonists, we selected three feature selection methods : T-test, Nearest Shrunken Centroids (NSC) and Support Vector Machine – Recursive Feature Elimination. These methods were tested on simulated and on real data. At the same time, a new methodology, MetRob, was developed in order to improve the robustness of these methods towards the technical variability of microarrays, as well as their reproducibility. This new methodology mainly improves the positive predictive value, which means the confidence in the results. The NSC method was found to be the most robust. The results of the association of MetRob and NSC were thus studied from a biological point of view.

Key words : Microarray, Feature selection, Data Mining, PPAR, Type 2 Diabetes

Remerciements

Je tiens à exprimer tout d'abord mes remerciements aux membres du jury, qui ont accepté d'évaluer mon travail de thèse.

Merci à M. Jean-Philippe Vert, Directeur du Centre for Computational Biology de Mines ParisTech, d'avoir accepté de présider le jury de cette thèse, et à MM. Pascal Barbry, directeur de l'IPMC, Sofia Antipolis, et Avner Bar-Hen, professeur à l'Université Paris Descartes, d'avoir accepté d'être les rapporteurs de ce manuscrit. Leurs remarques et suggestions lors de la lecture de mon rapport m'ont permis d'apporter des améliorations à la qualité de ce dernier.

Merci à Christian Saguez, pour avoir accepté de diriger cette thèse et pour la confiance et la liberté qu'il m'a accordées.

Je tiens à remercier aussi Françoise Xavier, dont le soutien et la présence constante, m'ont permis de mener ce travail à terme.

Merci également à Jean-Pierre Galizzi pour son implication dans ce travail, sa disponibilité et ses précieux conseils.

A Brian Lockhart, directeur de la division PPM à l'Institut de Recherches Servier, merci de m'avoir accueilli au sein de son équipe.

Je tiens à remercier l'ensemble de la division PPM de l'Institut de Recherches Servier et plus particulièrement Sophie G., Nolwen, Chantal, Sabrina et Sophie M. pour leurs conseils et leur accueil, ainsi que les membres de la division P03 qui ont réalisé les expériences sur les souris.

Un grand merci également à Sylvie et Corinne pour leur gentillesse et leur efficacité lors des difficultés administratives ou logistiques que j'ai rencontrées.

Je tiens enfin à remercier les amis, thésards ou non, qui m'ont aidé au cours des trois ans de cette thèse. Merci à mes cobureaux successifs, Marc, Takuya, Cédric et Véro, ainsi qu'aux équipes Masbio et Digiplante pour les discussions enrichissantes (professionnelles ou non…) et les sympathiques soirées pizza-jeux : Vincent, Qi Rui, Marlène, Xiu Juan, Qiongli, Zhong Ping, Benoît, Thomas, Fenni, Natacha, Guanghui, Frédérique, ... Merci également à mes amis de longue date qui ont supporté mes moments de doute : Elodie, Murielle, Maud et toutes les petites familles associées… Et merci à Mahendra pour m'avoir accompagnée pendant la dernière partie, la plus stressante (ah le résumé de rapport au téléphone…), de cette thèse.

Finalement j'adresse un grand merci à toute ma famille qui a toujours été présente lorsque j'en ai eu besoin, en particulier à mon frère, à mon père et à ma mère.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ACP	:	Analyse en Composantes Principales
ADN	:	Acide DesoxyriboNucléique
ARN	:	Acide Ribonucléique
ATP	:	Adénosine TriPhosphate
CoA	:	Coenzyme A
Cy3	:	Cyanine 3
Cy5	:	Cyanine 5
GA	:	Algorithmes Génétiques
ICED	:	Independently Consistent Expression Discriminator
IQR	:	Interquantile Range
KNN	:	K plus proches voisins
KNN-GA	:	Méthode de sélection de variables combinant algorithmes génétiques et K plus proches
LOOCV	:	Leave-One-Out-Cross-Validation
MAQC	:	Consortium pour l'évaluation de la reproductibilité inter et intra plate-formes des puces à
MetRob	:	Méthodologie robuste de sélection de gènes développée au cours de cette thèse
NE	:	Non Exprimé
NSC	:	Nearest Shrunken Centroids
NSCRob	:	MetRob couplée à Nearest Shrunken Centroids
PLS-DA	:	Partial Least Square Discriminant Analysis
PPAR	:	Peroxisome Proliferator-Activated Receptor
RFE	:	Recursive Feature Elimination
SAM	:	Significance Analysis of Microarrays
SCOMP	:	Composé développé par Servier, agoniste partiel de PPARγ
Se	:	Sensibilité
Séquences SSR	:	Séquences Statistiquement Significativement Régulées
Soleus	:	Muscle squelettique
Souris db/db	:	Modèle de souris diabétiques qui présentent une mutation sur le gène codant pour le récepteur à la leptine
Souris db+	:	Modèle de souris saines de la même souche que les souris db/db
Souris ob/ob	:	Modèle de souris diabétiques qui présentent une mutation sur le gène codant pour la leptine
Sp	:	Spécificité
SPPARM	:	Modulateur sélectif de PPAR
SPROD	:	Composé Servier testé sur les souris ob/ob, agoniste PPAR mixte α/γ
SVM	:	Support Vector Machines
SVM-RFE	:	Support Vector Machine – Recursive Feature Elimination
TAI	:	Tissu Adipeux Inguinal
VPN	:	Valeur Prédictive Négative
VPP	:	Valeur Prédictive Positive

Préambule

Ces travaux ont été effectués dans le cadre d'une bourse CIFRE en partenariat entre la division de Pharmacologie et Physiopathologie Moléculaires de l'Institut de Recherches Servier et le laboratoire de Mathématiques Appliquées aux Systèmes de l'Ecole Centrale Paris.

Le diabète de type 2, forme la plus courante de diabète, est une maladie métabolique chronique principalement caractérisée par une résistance à l'insuline, une hyperglycémie et une hyperlipidémie. De par son ampleur croissante, il est devenu un véritable enjeu de santé publique. Parmi les cibles thérapeutiques associées à cette maladie, les Peroxisome Proliferator-Activated Receptors (PPARs) sont des récepteurs nucléaires présents sous trois isotypes (α, β et γ). Certains médicaments actuellement sur le marché ciblent les formes α et γ de PPAR. Les agonistes de PPARα permettent de normaliser les triglycérides mais peuvent provoquer des désordres digestifs et musculaires. Les agonistes de PPARγ restaurent la sensibilité à l'insuline, mais sont susceptibles d'induire œdème et prise de poids. De nouvelles formes d'agonistes sont donc recherchées par l'industrie pharmaceutique : agonistes β, agonistes mixtes α/γ et agonistes γ partiels. C'est dans cette optique que l'Institut de Recherches Servier a développé un agoniste γ partiel (appelé SCOMP). L'objectif de cette thèse est d'étudier les différences entre un composé de référence, la rosiglitazone (agoniste γ), et le SCOMP (agoniste γ partiel) chez des souris diabétiques db/db. Pour ce faire, des analyses en puces à ADN ont été menées sur différents tissus, donnant accès à des mesures différentielles du transcriptome entre souris traitées et souris non traitées.

L'étude des différences entre rosiglitazone et SCOMP a été abordée comme un problème de sélection de variables (les gènes) dans le cadre de la discrimination entre souris traitées par la rosiglitazone et souris traitées par le SCOMP. De nombreuses méthodes permettant de répondre à cette question ont déjà été présentées dans la littérature. Ces méthodes ne prennent néanmoins pas en compte l'impact de la variabilité technique des puces à ADN sur leurs résultats. En outre, il est difficile de choisir un nombre de gènes optimal à conserver. Dans cette optique, trois méthodes de sélection de variables ont été choisies pour être testées et comparées, essentiellement en termes de robustesse : le T-test, la méthode de Nearest Shrunken Centroids et la méthode de Support Vector Machine - Recursive Feature Elimination. Les données ont été empiriquement perturbées de manière à reproduire la variabilité technique des puces à ADN. La robustesse a alors été évaluée comme la stabilité des listes de gènes générées par les méthodes quand elles étaient appliquées à ces jeux de données perturbées. Une nouvelle méthodologie, MetRob, a ensuite été développée afin

d'améliorer la robustesse et la reproductibilité des résultats d'une méthode de sélection de variables. Les deux problèmes précités ont été gérés conjointement en choisissant le nombre de gènes maximisant la robustesse.

Les trois méthodes de sélection de variables ainsi que MetRob ont été testées sur des jeux de données simulées et sur les jeux de données réelles de l'étude PPAR. Dans un premier temps, nous avons étudié différents aspects de l'efficacité de ces méthodes : robustesse, pertinence de la sélection, pouvoir discriminant des gènes et impact du nombre d'observations. Dans un second temps, nous avons cherché à valider les résultats de MetRob pour leur pertinence biologique. Les listes de gènes sélectionnées ont été analysées en termes de métabolisme, puis reliées à des paramètres biologiques mesurés au cours des expériences.

Ce manuscrit est organisé en 5 chapitres. Le **Chapitre 1** développe le contexte biologique et technologique de cette thèse. Diabète, PPAR et puces à ADN y sont abordés. La problématique au niveau mathématique, ainsi qu'un état de l'art des méthodes de sélection de variables existantes sont présentés dans le **Chapitre 2**. Le **Chapitre 3** porte sur la méthodologie de sélection de listes de gènes robustes développée au cours de cette thèse : mise en place, paramétrage et implémentation. Enfin, les **Chapitre 4** et **Chapitre 5** présentent les résultats obtenus en termes d'efficacité des méthodes et d'analyse biologique des listes de gènes sélectionnées.

Chapitre 1 :

Contexte biologique et technologique

1.1 Diabète

Le diabète est une maladie métabolique chronique qui affecte la régulation du taux de glucose dans le sang. Il est caractérisé par une glycémie élevée (glycémie à jeun supérieure à 1,26 g/L [1]) et peut provoquer hypertension, artérite, cécité, insuffisances rénales, etc. L'Organisation Mondiale de la Santé chiffre le nombre de diabétiques dans le monde à plus de 180 millions en 2008 et estime que ce nombre devrait plus que doubler d'ici 2030 [2]. Il existe plusieurs types de diabètes correspondant à des altérations différentes du fonctionnement normal du métabolisme. Le diabète de type 1, ou diabète insulino-dépendant, correspond à une déficience en insuline. Le diabète de type 2, ou diabète non insulino-dépendant, est caractérisé par une résistance à l'insuline. D'autres formes plus rares existent, comme le diabète gestationnel ou le diabète néonatal. Le diabète de type 2 est la forme prépondérante et représente environ 90% des diabétiques. Il induit un coût non négligeable pour les systèmes de santé nationaux (5 à 10% des budgets dans les pays développés) et s'étend à des classes de populations de plus en plus jeunes [3]. Cette maladie représente donc un véritable enjeu de santé publique qui mobilise l'industrie pharmaceutique.

1.1.1 Métabolisme d'un individu non diabétique

Après un repas (Figure 1.1.a), les aliments sont digérés et le glucose, les acides aminés et les lipides sont relâchés dans le sang [4]. La présence de glucose stimule la sécrétion d'insuline par les cellules β des îlots de Langerhans du pancréas. Cette insuline induit le stockage des sources d'énergie. Plus précisément, elle stimule l'absorption du glucose dans le foie et le muscle où il est stocké sous forme de glycogène ou métabolisé afin de fournir énergie et matières premières pour la synthèse des acides gras. Ces acides gras sont ensuite transformés en triglycérides puis transportés par le biais de lipoprotéines vers le tissu adipeux où ils sont stockés.

En période de jeûne (Figure 1.1.b), le taux de glucose dans le sang chute. La sécrétion d'insuline diminue et du glucagon est sécrété par les cellules α des îlots de Langerhans du pancréas. Le glucagon a pour cible principale le foie. Il y stimule la dégradation des stocks de glycogène et la néoglucogenèse (synthèse de glucose), permettant le relargage de glucose dans le sang. En parallèle, muscle et tissu adipeux diminuent leur absorption de glucose et le tissu adipeux libère des acides gras qui sont ensuite utilisés comme source d'énergie par le muscle et le foie. Le foie produit alors des corps cétoniques qui fournissent le cerveau en énergie en l'absence de glucose.

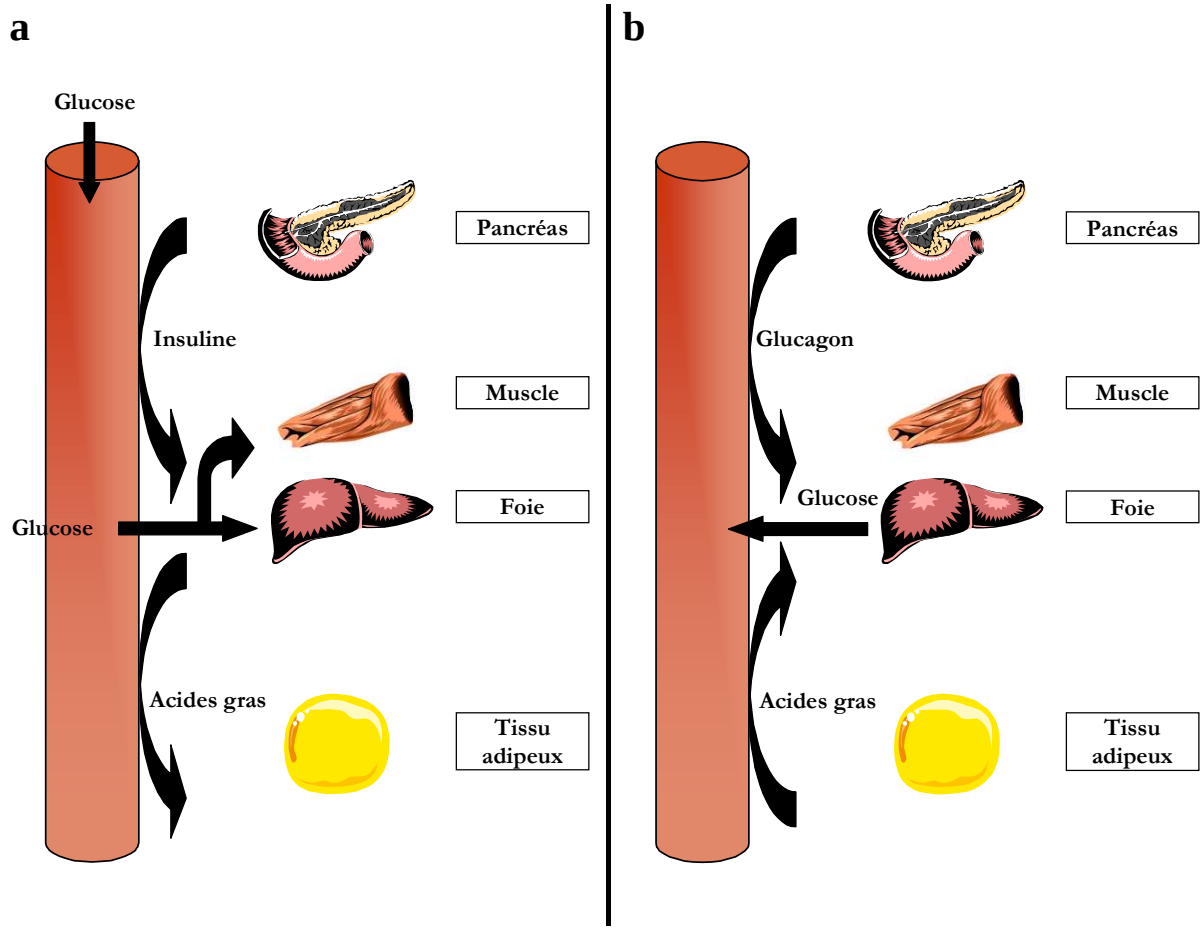

FIG. 1.1 : Métabolisme après un repas (a) et en période de jeûne (b)
a : La présence de glucose dans le sang stimule la sécrétion d'insuline par le pancréas. Le glucose est alors absorbé et traité par le muscle et le foie. Les acides gras produits par ces organes sont stockés dans le tissu adipeux sous forme de triglycérides. b : La faible quantité de glucose dans le sang stimule la sécrétion de glucagon par le pancréas. Le foie relargue alors du glucose dans le sang et le tissu adipeux libère des acides gras.

Chez les rongeurs, l'équilibre entre ces deux périodes est essentiellement maintenu par deux hormones : la leptine et le neuropeptide Y [5]. La leptine est produite par le tissu adipeux et inhibe la prise alimentaire. Le neuropeptide Y, produit par l'hypothalamus, a au contraire le rôle de stimulateur de la prise alimentaire. En fonctionnement normal, la concentration de neuropeptide Y augmente quand un animal a faim, stimulant la prise alimentaire, entraînant le stockage des nutriments aux dépens de leur utilisation et induisant la sécrétion de leptine qui joue un rôle de rétrocontrôle négatif.

1.1.2 Diabète de type 1

La sécrétion d'insuline provoque normalement l'absorption du glucose dans les organes consommateurs. Dans le cadre d'un diabète de type 1, le pancréas est incapable de sécréter suffisamment d'insuline. Cette défaillance est due à une maladie auto-immune qui détruit progressivement les cellules β des îlots de Langerhans du pancréas. Le glucose reste donc dans le sang et la glycémie augmente. Cette hyperglycémie peut avoir des conséquences microvasculaires (rétinopathie, insuffisance rénale chronique), macrovasculaires (hypertension, attaque cérébrale, artérite) et nerveuses (neuropathie). Ce type de diabète se déclare souvent très jeune et reflète une prédisposition d'origine génétique. Il ne représente que 10% des diabétiques et est essentiellement traité par injections d'insuline [6].

La recherche actuelle s'oriente vers de nouveaux modes d'administration de l'insuline moins contraignants que les injections (voie pulmonaire ou voie orale) [3]. Des greffes de tout ou partie du pancréas peuvent être effectuées, mais des études de thérapies par cellules souches embryonnaires sont actuellement menées pour pallier le manque de donneurs [7]. Les possibilités de régénération des cellules β en associant lutte contre le processus auto-immun et stimulation de la croissance cellulaire sont également étudiées [8].

1.1.3 Diabète de type 2

Le diabète de type 2 est la forme majoritaire du diabète et est caractérisé par une résistance des cellules cibles de l'insuline qui n'absorbent alors plus le glucose. Il est associé à une production insuffisante d'insuline et à une hyperlipidémie. La conséquence première est une augmentation de la glycémie avec les mêmes effets néfastes que pour le diabète de type 1. Le diabète de type 2 se développe classiquement chez les personnes âgées ou en surpoids. On observe néanmoins une évolution inquiétante du diabète de type 2 vers des populations plus jeunes, en association avec l'obésité infantile. Les causes de la maladie restent diverses et mal connues, mais âge, obésité, antécédents familiaux et origine ethnique sont des facteurs de risques. D'autre part, le diabète de type 2 est dans un premier temps asymptomatique et n'est donc souvent diagnostiqué qu'à l'apparition des premières complications, ce qui rend la prise en charge de la maladie très difficile [3][9].

Un mode de vie plus sain associant sport et alimentation équilibrée représente une part importante du traitement mais ce n'est pas toujours suffisant. Des médicaments régulant la glycémie après les repas sont alors administrés : sulfamidés hypoglycémiants, inhibiteurs des α-glucosidases, glinides. Ces hypoglycémiants ne sont néanmoins pas exempts d'effets secondaires (prise de poids, hypoglycémie, douleurs digestives) [10]. D'autres molécules dites sensibilisatrices à l'insuline ont également été développées : les thiazolidinediones (ex : rosiglitazone) et les biguanides. Œdème, prise de poids et mauvaise tolérance digestive peuvent cependant y être associés [11]. Enfin, les fibrates agissent sur l'hyperlipidémie, souvent liée au diabète de type 2, mais peuvent induire troubles digestifs ou musculaires [12]. En cas d'action insuffisante de ces traitements, le patient se voit proposer des injections d'insuline. D'autres types de molécules semblent prometteurs (mimétiques de GLP-1 et inhibiteurs de DPP IV), mais elles sont chères et leur sécurité à long terme n'est pas encore bien établie [13].

Du fait de l'ampleur considérable que prend le diabète de type 2 dans le monde, trouver de nouveaux médicaments aussi efficaces et présentant moins d'effets secondaires est donc devenu un enjeu primordial dans l'industrie pharmaceutique.

1.2 Modèles animaux du diabète de type 2

Les nouvelles molécules susceptibles de lutter contre le diabète de type 2 doivent auparavant être testées sur des modèles in vivo se rapprochant au mieux de la pathologie humaine. Il existe différents types de modèles animaux : modèles de diabète spontané (animaux génétiquement sélectionnés), modèles induits par un régime alimentaire, modèles induits par des produits chimiques, modèles induits par inoculation d'un virus ou modèles induits chirurgicalement. Le Tableau 1.1 résume les caractéristiques des principaux modèles de rongeurs disponibles [14][15][16][17]. Les modèles de diabète spontané comme les rats OLETF (Otsuka Long-Evans Tokushima Fatty Rats), les rats GK (Goto-Kakizaki Rats), les souris db/db, les rats ZDF (Zucker Diabetic Fatty Rats) et les souris ob/ob sont principalement utilisés dans la recherche pharmaceutique [17]. Dans le cadre de notre étude, le modèle de souris db/db a été choisi.

Les souris db/db présentent une mutation du gène codant pour les récepteurs de la leptine [15]. La leptine est une hormone inhibant la prise alimentaire chez le rongeur et jouant un rôle de rétrocontrôle négatif vis à vis du neuropeptide Y qui stimule la prise alimentaire. Chez les souris db/db, la leptine n'est plus reconnue par ses organes cibles. Le rétrocontrôle négatif n'existe plus et l'animal devient hyperphagique, provoquant ainsi une obésité [5]. Les souris db/db développent les symptômes d'un diabète de type 2 dès 6 semaines. Elles sont hyperglycémiques, insulino-résistantes et transitoirement hyperinsulinémiques. Ces souris souffrent également de dyslipidémie avec des niveaux élevés de cholestérol [18]. Les souris db/db ont été largement utilisées pour l'étude du diabète de type 2 et le test de molécules sensibilisatrices à l'insuline.

	Type de modèle	Obésité	Insulino-résistance	Hyper-glycémie	Hyper-insulinémie
Souris ob/ob	Diabète spontané (mutation dans le gène de la leptine)	+	+	+ (transitoire)	+
Souris db/db	Diabète spontané (mutation dans le gène du récepteur à la leptine)	+	+	+	+ (transitoire)
Souris NZO	Diabète spontané	+	+	+ (faible)	+
Souris KK/Ay (Yellow)	Diabète spontané	+	+	+	+
Rat OLETF	Diabète spontané	+ (faible)	+	+	+
Rat GK	Diabète spontané	-	+	+ (modérée)	-
Rat ZFR	Diabète spontané (mutation dans le gène fa)	+	+	+ (faible)	+
Rat ZDR	Diabète spontané (souche des rats ZFR sélectionnés pour leur hyperglycémie)	+	+	+	+ (transitoire)
Souris Spiny	Régime alimentaire (régime sucrose)	-	+ (faible)	-	+ (faible)
Souris Spiny	Régime alimentaire (régime lipides)	+	+ (faible)	+ (faible et transitoire)	+ (faible et transitoire)
Rat des Sables	Régime alimentaire (régime standard de laboratoire)	+	+	+	+

TAB. 1.1 : Caractéristiques des principaux modèles animaux de diabète de type 2
Les modèles de diabète spontané dont les caractéristiques ne sont pas détaillées sont des modèles polygéniques obtenus par croisements sélectifs. Souris NZO : New Zealand Obese Mouse, Rat OLETF : Otsuka Long-Evans Tokushima Fatty Rat, Rat GK : Goto-Kakizaki Rat, Rat ZFR : Zucker Fatty Rat, Rat ZDR : Zucker Diabetic Fatty Rat.

1.3 Peroxisome Proliferator-Activated Receptors

1.3.1 PPAR, un récepteur nucléaire impliqué dans le métabolisme

Les Peroxisome Proliferator-Activated Receptors (PPARs) sont des récepteurs nucléaires, cibles de plusieurs médicaments contre le diabète de type 2. Un récepteur nucléaire est un facteur de transcription qui, activé par un ligand, induit la transcription de gènes cibles (Figure 1.2). Lorsqu'il est activé, PPAR recrute différents cofacteurs, s'hétérodimérise avec le récepteur rétinoïde RXR et se fixe sur les éléments de réponse des promoteurs de ses gènes cibles. PPAR est activé par plusieurs catégories de lipides et régule essentiellement des gènes impliqués dans le métabolisme des acides gras et dans la réponse inflammatoire. Il existe 3 isotypes de PPAR : α, β (ou δ) et γ.

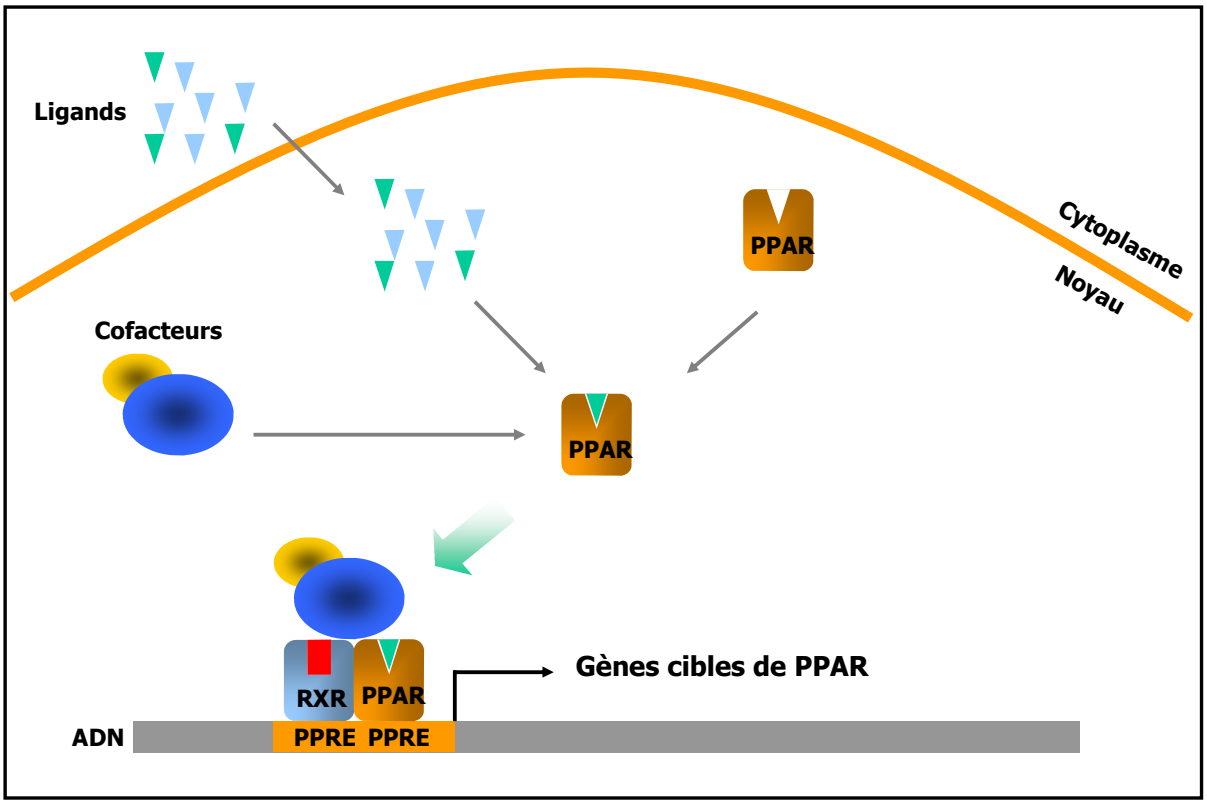

FIG. 1.2 : Activation des récepteurs nucléaires de type PPAR
Après activation par des ligands et en présence de cofacteurs, PPAR s'hétérodimérise avec RXR, puis se fixe sur les éléments de réponses (PPRE) des promoteurs de ses gènes cibles pour en activer la transcription.

PPARα est surtout exprimé dans le foie. Il régule des gènes impliqués dans la lipolyse des lipoprotéines de très basse densité (VLDL), ainsi que dans la capture des acides gras et leur dégradation intracellulaire. Il intervient également dans la fonction vasculaire et l'athérogenèse. PPARγ est principalement exprimé dans le tissu adipeux. Il régule des gènes impliqués dans la différentiation adipocytaire et dans l'accumulation des lipides par les adipocytes. Son activation contribuerait ainsi à réduire la quantité d'acides gras et de triglycérides circulants. PPARγ régule également l'expression de l'adiponectine, intervenant par ce biais dans les processus inflammatoires et athérosclérotiques. PPARβ est exprimé de manière ubiquitaire dans l'organisme. Son rôle n'est pas encore très bien connu [19][20].

1.3.2 Des agonistes PPAR contre le diabète de type 2

L'impact de l'activation de PPAR sur de nombreux gènes liés au métabolisme a conduit à le considérer comme une cible pharmacologique contre le diabète de type 2 [21]. Plusieurs médicaments actuellement sur le marché sont des agonistes PPAR. Les thiazolidinediones sont des molécules sensibilisatrices à l'insuline, agonistes de PPARγ, à la famille desquelles appartiennent la rosiglitazone et la pioglitazone. Les thiazolidinediones restaurent la sensibilité à l'insuline chez les patients diabétiques mais leur action sur les lipides reste limitée et elles peuvent induire des effets secondaires comme l'œdème, la prise de poids et l'ostéoporose [11][22]. Les hypolipidémiants tels que les fibrates n'ont pas, à l'origine, été développés pour lutter contre le diabète de type 2. Néanmoins, le diabète de type 2 étant souvent associé à une hyperlipidémie, ces agonistes de PPARα se sont révélés utiles contre la pathologie. Les fibrates normalisent les triglycérides et augmentent le taux de lipoprotéines de haute densité (HDL) mais ils sont administrés à fortes doses et sont susceptibles d'induire troubles digestifs et musculaires [12].

	Effet sur la glycémie	**Effet sur les lipides**	**Effets secondaires possibles**
Agonistes PPARγ	Restaurent la sensibilité à l'insuline	Limité	Oedème Prise de poids Ostéoporose
Agonistes PPARα	Non	Normalisent les triglycérides Augmentent les HDL	Troubles digestifs Atteintes musculaires
Agonistes mixtes PPARα/γ	Restaurent la sensibilité à l'insuline	Normalisent les triglycérides Augmentent les HDL	Accidents cardiovasculaires (muraglitazar)
Agonistes SPPARMγ	Restaurent la sensibilité à l'insuline	Limité	NA

TAB. 1.2 : Tableau récapitulatif des classes de médicaments agonistes PPAR
SPPARM : agoniste modulateur sélectif de PPAR

Des agonistes mixtes α/γ de PPAR ont également été développés dans le but de combiner sensibilisation à l'insuline et action sur les lipides. Jusqu'à présent, aucune de ces molécules n'a pu être mise sur le marché : le muraglitazar a été arrêté en cours de développement car il engendrait des accidents cardiovasculaires [23]. Cette approche demeure néanmoins intéressante d'un point de vue stratégique. En parallèle, une recherche d'agonistes modulateurs sélectifs de PPARγ (SPPARM) ou agonistes γ partiels est également d'actualité [24]. Ce type de molécules se fixe à PPARγ mais induit l'expression d'un spectre réduit de gènes par rapport à un agoniste plein selon l'organe considéré. Une telle approche cherche à conserver les propriétés antidiabétiques de la rosiglitazone tout en réduisant les effets secondaires. Le Tableau 1.2 récapitule ces différentes classes de médicaments agonistes PPAR.

Les travaux de cette thèse ont essentiellement porté sur l'étude des différences entre la rosiglitazone (agoniste PPARγ de référence) à un agoniste SPPARM synthétisé par la société Servier, dénoté par la suite SCOMP. Le SCOMP est un agoniste de PPARγ partiel dont l'activation in vitro est à 20% de celle d'un agoniste γ complet dans le tissu adipeux, à 90% dans le foie et qui est également capable d'activer PPARα. Il a été caractérisé dans le Département d'Athérosclérose, INSERM U545, Institut Pasteur de Lille (P. Lefebvre). Les résultats de cette caractérisation sont donnés en Annexe A.

1.4 Puces à ADN

Lors de l'évaluation d'une nouvelle molécule thérapeutique chez un modèle animal, la mesure de différents paramètres biologiques (ex : glycémie dans le cadre du diabète de type 2) permet d'étudier l'efficacité d'un composé. Ce type de mesure ne permet néanmoins pas de comprendre le mode d'action de la molécule testée. Dans le cadre de la recherche d'agonistes SPPARM, on souhaite essentiellement montrer que le mode d'action du composé testé est différent de celui d'un agoniste PPARγ complet. C'est dans cette optique que des mesures d'expression des gènes sur puces à ADN ont été réalisées pour l'ensemble du transcriptome de la souris.

1.4.1 Principe des puces à ADN

Les puces à ADN [25] permettent de mesurer simultanément le niveau d'expression de plusieurs milliers de gènes dans un échantillon biologique, c'est-à-dire les ARNs messagers. Ce sont des lames sur lesquelles sont fixées des sondes d'ADN. Les ARNs messagers d'un échantillon biologique, préalablement extraits, amplifiés et marqués, sont hybridés sur la lame. Chaque séquence d'ARN messager se fixe alors sur la sonde complémentaire correspondante. La lame est ensuite lue par un scanner et les résultats sont analysés pour

fournir le niveau d'expression de chaque gène d'intérêt. Ces informations sont habituellement exploitées par la comparaison des profils d'expression géniques entre différentes conditions biologiques.

Il existe différents types de lames et de protocoles. Une lame de puce à ADN est un support solide de quelques centimètres carrés sur laquelle sont fixées des sondes pouvant être de deux natures différentes. Ce sont soit de longues séquences d'ADN complémentaire (500 à 5000 bases) qui ont été déposées sur la lame mais sont de moins en moins utilisées, soit des oligonucléotides plus courts (20 à 80 bases) déposés sur la lame ou synthétisés in situ. L'approche de mesure de l'expression des gènes en puces à ADN peut être directe ou différentielle. Dans la version directe, un seul échantillon biologique est hybridé sur la lame et mesuré. Les résultats obtenus ne sont pas quantitatifs car toutes les sondes ne fixent pas de la même manière leur séquence complémentaire. L'analyse différentielle consiste à hybrider sur la même lame un échantillon contrôle de référence mélangé à l'échantillon à tester, chacun marqué par un fluorochrome différent. Ce type d'approche permet de s'affranchir des différences d'affinité des sondes.

1.4.2 Technologie Agilent utilisée

1.4.2.1 Caractéristiques des lames Agilent

Les puces à ADN utilisées pour les expériences étudiées sont des puces à oligonucléotides (60 bases) commercialisées par la société Agilent. Les lames sont des lames de souris 4*44k [26], c'est-à-dire que chaque lame comporte environ 4*44000 sondes permettant de mesurer simultanément l'expression de tout le génome de la souris pour 4 échantillons biologiques (pattern ID 014868).

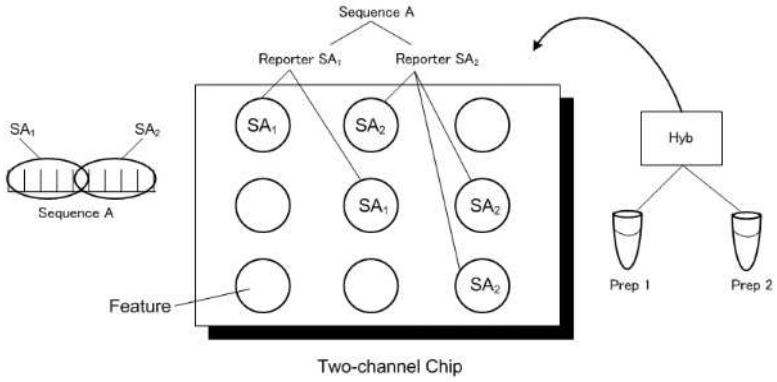

FIG. 1.3 : Différents niveaux de reconnaissance des séquences d'ARN [27]
Un feature est un spot sur lequel sont adsorbées plusieurs sondes identiques (reporter). Le même reporter peut être adsorbé sur différents features. Plusieurs reporters différents peuvent reconnaître la même séquence d'ARN.

Chaque spot de la lame, ou feature, est constitué d'un ensemble de sondes identiques (reporters ou oligonucléotides) qui permettent le dosage des ARNs complémentaires. Plusieurs features peuvent correspondre au même reporter, plusieurs reporters différents peuvent doser la même séquence d'ARN et plusieurs séquences peuvent correspondre au même gène (transcrits différents). Ceci est illustré sur la Figure 1.3.

1.4.2.2 Analyse différentielle

Une analyse différentielle est utilisée et le protocole expérimental représenté sur la Figure 1.4 est le suivant. Les ARNs sont extraits pour les animaux témoins et traités. Les échantillons d'ARN des animaux témoins sont mélangés pour constituer un échantillon contrôle. Ces ARNs sont amplifiés et marqués à la Cyanine 3 (cf Annexe A pour le détail de l'amplification et du marquage). Les échantillons d'ARN des animaux traités sont amplifiés et marqués à la Cyanine 5. Chaque échantillon d'un animal traité est mélangé avec l'échantillon contrôle, puis hybridé sur une puce à oligonucléotides Agilent.

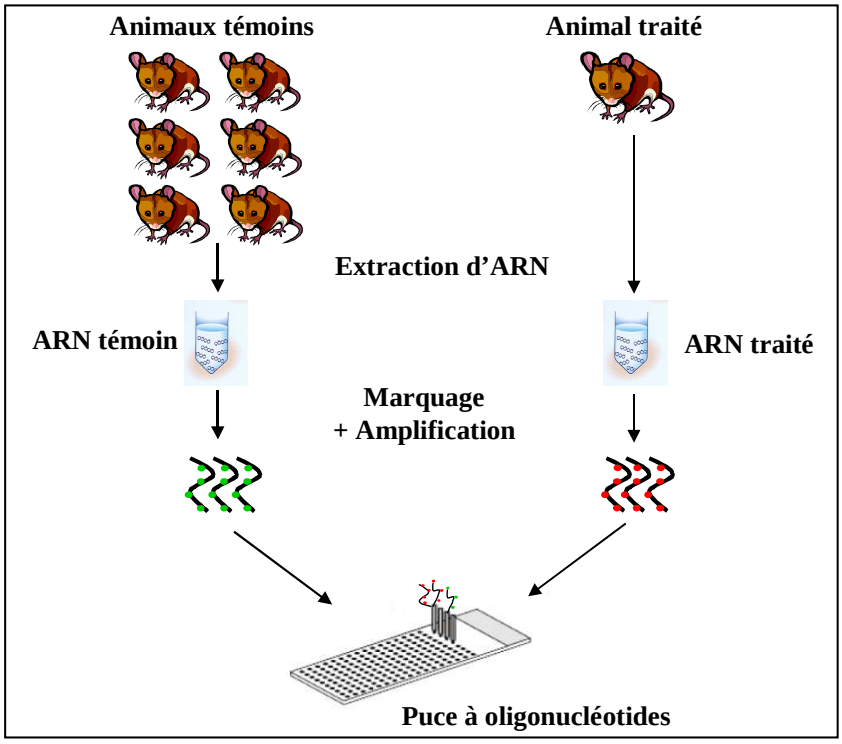

FIG. 1.4 : Protocole expérimental d'analyse différentielle
Les ARNs des animaux témoins sont regroupés et marqués à la Cyanine 3. L'ARN d'un animal traité est marqué à la Cyanine 5. Les deux échantillons sont mélangés, puis hybridés sur une puce à oligonucléotides Agilent.

Après l'hybridation, les lames sont lavées puis lues par un scanner. La Cyanine 3 (Cy3) apparaît en vert et la Cyanine 5 (Cy5) en rouge. La couleur de chaque spot informe sur la différence d'expression entre témoin et traité pour le feature correspondant (Figure 1.5) :

- spot apparaissant en noir : pas d'expression
- spot apparaissant en jaune : expression similaire chez le témoin et le traité
- spot apparaissant en vert (Cy3) : sous-expression chez le traité par rapport au témoin
- spot apparaissant en rouge (Cy5) : sur-expression chez le traité par rapport au témoin

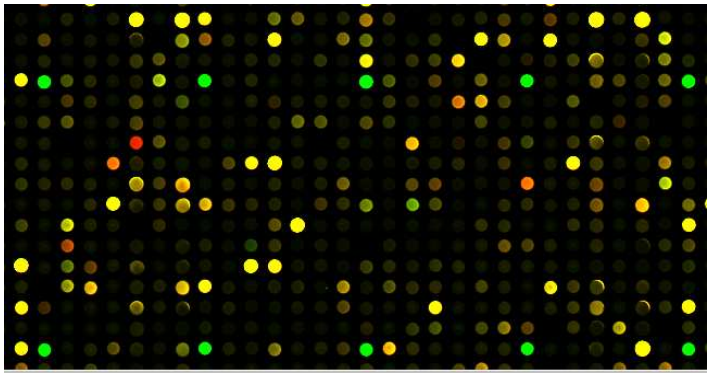

FIG. 1.5 : Image scannée d'une puce à ADN

Un logiciel de traitement d'image (Feature Extraction, Agilent) extrait les données à partir de l'image scannée : intensités des canaux Cy3 et Cy5, mesure d'erreur, rapport d'expression entre témoin et traité, significativité de la différence d'expression, … Différents biais expérimentaux ou techniques peuvent intervenir au cours du processus : petites différences dans les quantités initiales d'ARN, déséquilibre entre les fluorochromes en terme de propriétés physiques (capacité d'incorporation à l'ARN, sensibilité à la lumière, demi-vie), limites techniques du scanner, hybridation non uniforme, marques de lavage, … C'est pourquoi le logiciel d'extraction procède à certains ajustements par rapport aux intensités brutes relevées.

1.4.2.3 Procédure de dye-swap

Une procédure appelée dye-swap (inversion des colorants : Figure 1.6) est utilisée afin de corriger en partie le biais lié aux différences entre les deux fluorochromes (Cy3 et Cy5). Chaque hybridation est réalisée deux fois : la première en appliquant les marquages cités précédemment et la seconde en inversant les fluorochromes entre échantillons témoins et traités. Les résultats sont ensuite combinés à l'aide d'un autre logiciel (Rosetta Resolver) pour obtenir un seul jeu de données par animal.

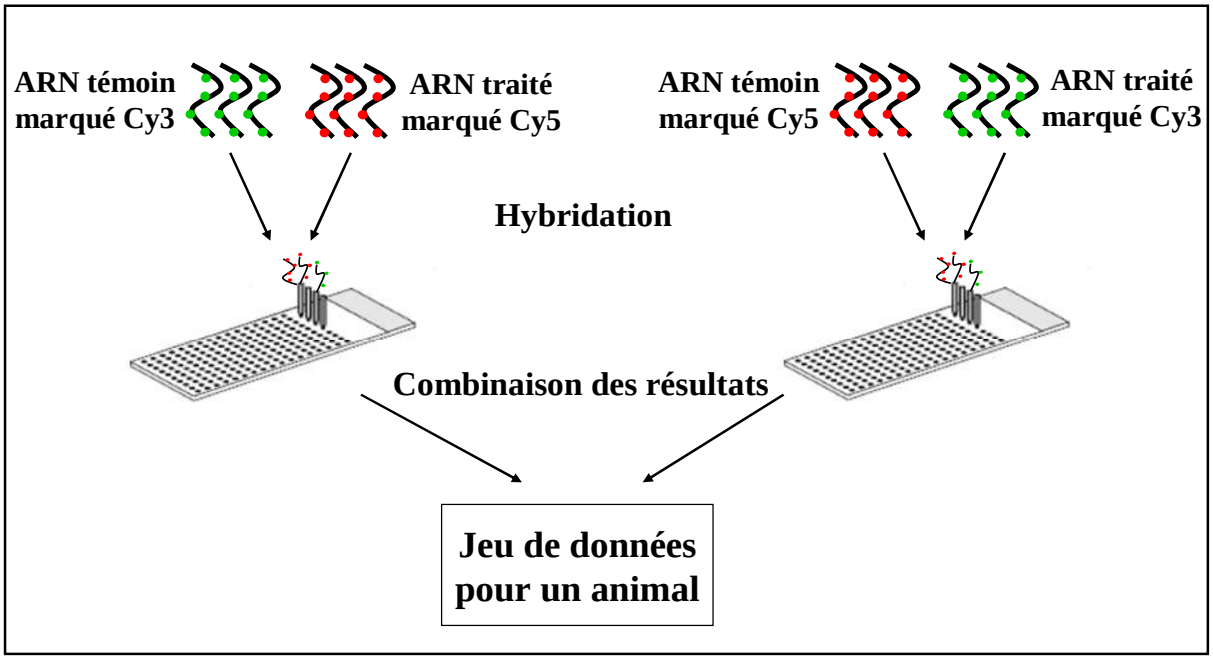

FIG. 1.6 : Procédure de dye-swap
Cy3 : Cyanine 3, Cy5 : Cyanine 5

1.5 Logiciels de traitement des données et d'analyse

Le résultat d'une analyse en puces à ADN est une image dans laquelle chaque spot coloré représente une sonde. Différents logiciels permettent d'exploiter ces résultats. Le logiciel Feature Extraction précité permet d'avoir accès à des valeurs de ratio d'expression pour chaque sonde et corrige une quantité importante de biais liés à l'expérience. Le logiciel Rosetta Resolver permet de combiner les données obtenues par animal, par produit, par expérience. Il permet également différents types d'analyses statistiques sur les résultats. Le logiciel Ingenuity Pathway Analysis comprend quant à lui une base de données consolidée permettant une analyse fonctionnelle de listes de gènes régulés.

1.5.1 Logiciel Feature Extraction

Feature Extraction [28] est un logiciel développé par Agilent qui permet d'obtenir des informations quantitatives à partir de l'image scannée d'une puce à ADN. La version 9.5 a été utilisée. Le détail du protocole appliqué dans notre étude est donné en Annexe B. Seules les grandes lignes du traitement sont reprises ici.

Le logiciel identifie tout d'abord la position des spots (features), puis il détermine pour chacun une zone centrale correspondant au signal et une zone périphérique correspondant au bruit de fond. Les pixels ayant des valeurs extrêmes ne sont pas pris en compte dans les calculs. Si plus de 50% des pixels d'un spot sont saturés pour les deux couleurs (vert pour Cy3 et rouge pour Cy5), ce spot sera considéré comme une valeur manquante. Ce traitement permet de gérer les limites techniques du scanner. Plusieurs tests sont ensuite effectués pour détecter les spots anormaux : spots non uniformes pour le signal ou le bruit de fond ou spots ayant des valeurs extrêmes dans une population de réplicats. Les spots non uniformes pour le signal seront également considérés comme des valeurs manquantes. Ce type de spots correspond généralement à des marques de lavage. Les spots ayant un signal significativement différent du bruit de fond sont également signalés.

Le signal est corrigé par soustraction d'un effet spatial additif pouvant être lié à une hybridation non uniforme et par division d'un effet spatial multiplicatif potentiellement dû à des différences de vitesses de réaction entre le centre de la lame et la périphérie. Une normalisation des intensités est ensuite réalisée afin de corriger une partie du biais lié aux différences entre les deux fluorochromes utilisés. Ce biais est connu pour dépendre entre autres de l'intensité et de la position sur la lame. La méthode utilisée est « linear and lowess ». Une régression linéaire est tout d'abord effectuée sur chaque canal afin de ramener la moyenne géométrique des intensités Cy3 et Cy5 à 1000. Le biais restant est estimé par régression locale (Locally weighted scatterplot smoothing : voir Annexe B) sur le MA-Plot (Figure 1.7 : graphe représentant M, le logarithme du rapport des intensités, en fonction de A, la moyenne logarithmique des intensités). Une régression locale est effectuée afin de trouver la tendance centrale des données, puis les données sont ajustées de manière à centrer le graphe en 0. Les biais liés aux fluorochromes indépendants de l'intensité et dépendants de l'intensité sont ainsi corrigés.

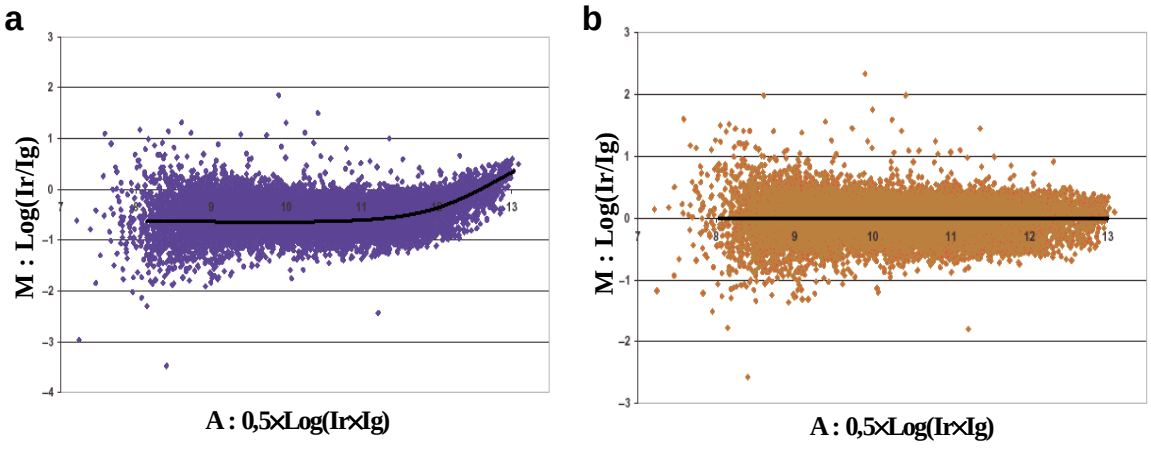

FIG. 1.7 : MA-plot représentant le logarithme du rapport des intensités en fonction de la moyenne logarithmique des intensités [29]

a : graphe avant normalisation lowess, b : graphe après normalisation lowess
Ig=Intensité du canal Cy3 (vert), Ir=Intensité du canal Cy5 (rouge)

Après normalisation, le log ratio de chaque spot est calculé : $\text{LogRatio} = \log(\frac{I_r}{I_g})$, où I_r est l'intensité du canal Cy5 (rouge) et I_g l'intensité du canal Cy3 (vert). Enfin, la significativité de la régulation des gènes entre les deux échantillons biologiques hybridés est déterminée pour chaque spot par l'utilisation du plus conservatif des deux modèles d'erreurs suivants :

- Propagated error model (modèle d'erreur propagée) : La dispersion des observations est estimée à partir de la dispersion des pixels.
- Universal error model (modèle d'erreur universel) : La dispersion des observations est estimée à partir d'un modèle supposant un effet additif et un effet multiplicatif par rapport à l'intensité du signal. L'erreur sur l'intensité I est donnée par $\sigma^2 = A^2 + M^2 \times I^2$. Le paramètre M^2 a été estimé à partir d'expériences réalisées par Agilent. Le paramètre A^2 est estimé automatiquement sur chaque lame.

Le modèle fournissant la plus grande erreur est utilisé pour calculer une p-value, c'est-à-dire la probabilité d'obtenir le même log ratio sous l'hypothèse nulle de non-régulation du spot. Le logiciel fournit donc notamment pour chaque spot deux valeurs d'intensités normalisées (Cy5 en rouge et Cy3 en vert), le rapport d'expression entre les deux canaux, une erreur sur ce rapport d'expression et une p-value informant sur la significativité de la régulation entre les échantillons. Ces résultats prennent en compte la gestion de plusieurs biais expérimentaux ou techniques. Les limites de mesure du scanner et la possibilité de marques de lavage sont gérées par le marquage des spots saturés ou non uniformes comme données manquantes. La correction de l'effet spatial du signal permet entre autres de compenser une hybridation non uniforme. Enfin, le déséquilibre entre les fluorochromes est traité à deux niveaux. La normalisation des intensités permet de gérer le biais dépendant de l'intensité et la procédure de dye-swap citée précédemment permet de gérer le biais dépendant de la séquence d'ARN considérée.

1.5.2 Logiciel Rosetta Resolver

Les données fournies par Feature Extraction sont ensuite importées dans le logiciel Rosetta Resolver [27] pour y être visualisées et analysées. Resolver permet d'observer les résultats pour une même lame au niveau des features, des reporters, des séquences ou des gènes et de combiner plusieurs lames correspondant au même traitement afin de rendre les résultats plus robustes (par exemple combinaison par dye-swap pour un même animal). Il fournit également des outils de visualisation (diagrammes de Venn, graphes de dispersion des intensités ou des log ratios, etc.) et des outils statistiques (analyse en composantes principales, analyse des variances, classification, etc.) permettant une première analyse des résultats. Resolver permet également de créer des ensembles de séquences prédéfinis qui peuvent servir de base pour les analyses statistiques (par exemple classification non supervisée à partir des séquences sélectionnées).

La méthode « Combining » permet de fusionner des lames correspondant aux mêmes conditions de traitement. Pour n lames à combiner, le nouveau log ratio est calculé par une moyenne pondérée par l'erreur sur le log ratio de sorte que les lames ayant une erreur plus importante influent moins sur le résultat final. Le modèle d'erreur développé par Rosetta a inspiré celui utilisé par Feature Extraction sans être totalement équivalent. Le détail de l'erreur utilisée est donné en Annexe B. Cette méthode est très utile pour définir des expériences où les données peuvent être regroupées par animal ou par traitement. Ces expériences peuvent ensuite être directement utilisées pour les analyses statistiques. La méthode « Squeezing » permet de passer des features aux reporters, des reporters aux séquences et des séquences aux gènes. Le principe est globalement le même que pour le « Combining ».

Il faut néanmoins noter que les informations obtenues à partir de Feature Extraction ne sont pas toutes reprises dans Resolver par la suite. Par exemple, la significativité du signal par rapport au bruit de fond pour chaque spot n'est pas importée dans Resolver.

1.5.3 Logiciel Ingenuity Pathway Analysis

Ingenuity Pathway Analysis (IPA) [30] est un logiciel d'analyse biologique, complémentaire de Resolver, qui comprend une base de données consolidée recensant les interactions décrites dans la littérature entre les gènes (régulations géniques, interactions protéiques, …), ainsi que les fonctions des gènes quand elles sont connues. Il permet de placer les séquences qui ont été considérées comme différentiellement régulées entre deux conditions biologiques dans des mécanismes biologiques, des voies de signalisation et des fonctions connues, répondant ainsi à diverses questions sur les gènes d'intérêt :

- Sont-ils reliés entre eux dans des réseaux communs ?
- Remplissent-ils des fonctions similaires ?
- Interviennent-ils dans la même voie métabolique ou de signalisation ?

IPA est compatible avec de nombreuses plate-formes d'analyse à haut débit de type puces à ADN. Il reconnaît entre autres les codes des séquences provenant de Resolver. Néanmoins les mises à jour des bases de données de Resolver et d'Ingenuity n'étant pas synchronisées, il peut survenir des différences d'annotations entre les deux logiciels.

1.6 Protocoles expérimentaux

Les travaux de cette thèse ont essentiellement porté sur l'étude de deux agonistes PPAR cités précédemment : la rosiglitazone (agoniste PPARγ) et le SCOMP (agoniste SPPARM). Ces deux molécules ont été testées chez des souris diabétiques db/db et les échantillons biologiques obtenus ont été hybridés sur des puces à ADN. Le protocole expérimental de cette étude, ainsi que le protocole d'une étude de variabilité technique des puces à ADN sont présentés ci-après. D'autres molécules agonistes PPAR ont également été testées sur un autre modèle de souris mais les résultats obtenus ont essentiellement été utilisés pour paramétrer les différentes méthodes mathématiques mises en place. Ces expériences sont détaillées en Annexe A.

1.6.1 Etude PPAR

Un effet dose pour la rosiglitazone et le SCOMP a été réalisé sur des souris mâles db/db âgées de 8 semaines (Figure 1.8). Les souris ont aléatoirement été divisées en huit groupes (n=9 par groupe) :

- souris db/db diabétiques non traitées (considéré comme le groupe témoin)
- souris non diabétiques db+
- souris db/db diabétiques traitées avec des doses de 28, 84 ou 280 µmol/kg de rosiglitazone (agoniste PPARγ)
- souris diabétiques db/db traitées avec des doses de 28, 84 ou 280 µmol/kg de SCOMP, agoniste modulateur sélectif de PPARγ

Les souris ont été euthanasiées après 18 jours de traitement. Le foie, le tissu adipeux inguinal (TAI) et le muscle squelettique (Soleus) ont été prélevés et coupés en deux. Six souris par groupe ont été choisies en fonction de la qualité de leurs ARNs messagers pour une analyse différentielle en puces à ADN de la première moitié des organes, le groupe témoin étant constitué des souris db/db non traitées. L'autre moitié des organes a été envoyée au centre Servier d'Orléans pour une analyse lipidomique. Des mesures biologiques ont également été réalisées durant l'expérience : poids des animaux, poids des organes, glycémie et taux d'hémoglobine glycquée (HbA1c). On peut noter que les foies de toutes les souris ont été analysés en puce à ADN pour permettre une étude de l'impact du nombre d'animaux sur les résultats. Néanmoins, sauf précision contraire, les groupes de souris pour un traitement seront considérés par la suite de taille égale à 6.

Le modèle de souris db/db a été utilisé car il est représentatif de l'insulino-résistance associée au diabète de type 2. On peut néanmoins noter que les doses administrées dans ces expériences sont assez élevées. En effet, la dose la plus faible de 28 µmol/kg (soit 10mg/kg) représente déjà une dose pharmacologique classique chez les rongeurs [31].

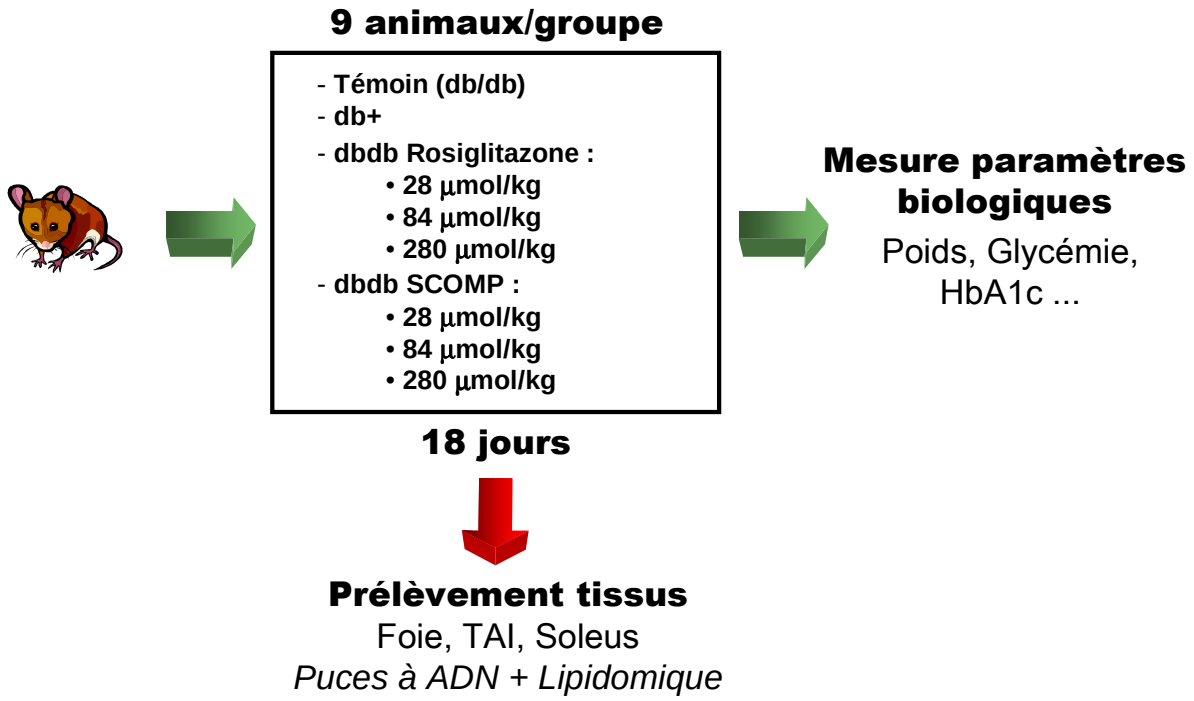

FIG. 1.8 : Protocole expérimental pour l'étude d'agonistes PPAR chez des souris db/db
TAI = Tissu adipeux inguinal, Soleus= Muscle squelettique, HbA1c= Hémoglobine glycquée

1.6.2 Etude de variabilité technique

Une étude de la variabilité technique des puces à ADN a été réalisée afin de pouvoir la quantifier et la reproduire. L'objectif est de tester la robustesse des méthodes mathématiques mises en place vis à vis de cette variabilité.

Deux échantillons biologiques de foie obtenus par le protocole expérimental cité précédemment ont été utilisés pour cette étude (rosiglitazone et SCOMP à la dose de 84 µmol/kg). Pour chaque échantillon, six réplicats techniques ont été obtenus. Le marquage a été réalisé à trois dates différentes et l'hybridation à deux dates différentes, puis tous les couples possibles ont été considérés (Figure 1.9). Une procédure de dye-swap a été utilisée (définition en 1.4.2).

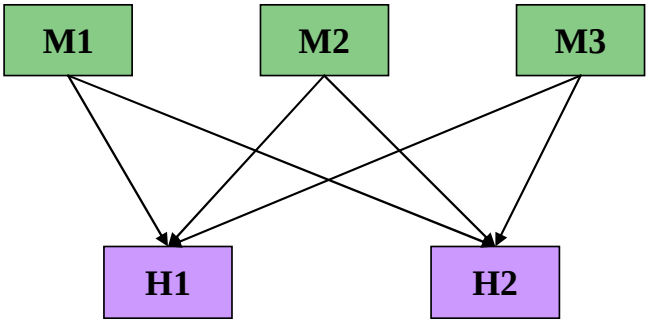

FIG. 1.9 : Protocole de l'étude de variabilité technique
M = marquage, H = hybridation

1.7 Problématique biologique

L'objectif de cette étude est de montrer que le mode d'action du composé SPPARM Servier, le SCOMP, est différent de celui d'un agoniste PPARγ de référence, la rosiglitazone. On espère obtenir avec ce nouveau composé une efficacité similaire à la rosiglitazone mais moins d'effets secondaires. C'est dans cette optique que des analyses en puces à ADN ont été menées. On souhaite, en pratique, rechercher les gènes qui expliquent au mieux les différences entre les deux produits, afin de mieux comprendre les différences observées biologiquement (mesure de paramètres biologiques et lipidomique).

Dans cette optique, les deux composés ont été comparés par dose et par organe, conduisant à 9 jeux de données présentés dans le Tableau 1.3. Chaque jeu de données consiste en l'union des données d'expression obtenues pour les souris traitées à la rosiglitazone et au SCOMP. Les variables considérées sont les reporters définis précédemment : oligonucléotides reconnaissant une petite séquence d'ARN. On parlera par la suite de séquences pour simplifier la lecture. Pour chaque animal et chaque séquence, on dispose du log ratio, de la p-value et des intensités normalisées Cy3 et Cy5. Des analyses comparatives entre souris diabétiques traitées (par la rosiglitazone ou le SCOMP) et souris non diabétiques db+ ont également été menées en parallèle afin de mieux comprendre le contexte physiopathologique.

	Muscle squelettique	**Foie**	**Tissu adipeux inguinal**
28 µmol/kg	RS_Soleus_dose1	RS_Foie_dose1	RS_TAI_dose1
84 µmol/kg	RS_Soleus_dose2	RS_Foie_dose2	RS_TAI_dose2
280 µmol/kg	RS_Soleus_dose3	RS_Foie_dose3	RS_TAI_dose3

TAB. 1.3 : Jeux de données pour l'étude des différences entre rosiglitazone et SCOMP

Chapitre 2 :

Problématique mathématique

Afin de répondre à cette problématique biologique d'étude des différences entre la rosiglitazone et le composé SCOMP, une approche de type « Data Mining » semble la plus adaptée. En effet, la recherche de gènes discriminant au mieux le type d'agoniste PPAR administré est un problème de sélection de variables dans le cadre de la discrimination entre souris traitées par la rosiglitazone et souris traitées par le SCOMP. La discrimination consiste à classer des individus par classes (le traitement dans notre étude) selon un certain nombre de caractéristiques (les expressions des gènes). La sélection de variables consiste à choisir les caractéristiques pertinentes permettant une classification efficace et limitant le risque d'erreurs. L'objectif est de construire une fonction qui associe à ces caractéristiques la classe de l'individu. Pour cela, les exemples d'individus disponibles sont utilisés. Le principal problème dans notre cas est le faible nombre d'individus (12) comparé au nombre de caractéristiques (plusieurs milliers), ce qui limite les possibilités d'application de méthodes classiques. Dans ce chapitre, le problème est formulé de manière plus rigoureuse d'un point de vue statistique, puis les solutions existantes sont passées en revue. La majorité des concepts statistiques présentés ici proviennent du cours d'analyse de données du Master 2 de probabilités et statistiques d'Orsay [32] et peuvent être retrouvés chez Hastie et al. [33].

2.1 Formalisation du problème

2.1.1 Discrimination et sélection de variables

Une méthode de discrimination permet de définir des règles de classification à partir de l'observation d'un jeu de données. Soient n individus (ou observations) sur lesquels sont mesurées p variables quantitatives $X = (X_1 , \dots , X_p)$. On note x_{ik} la réalisation de la variable i pour l'individu k et $x_k = (x_{1k} , \dots , x_{pk})$ le vecteur des variables pour l'individu k. Soit Y la variable aléatoire correspondant à la classe des observations : y_k est la classe de l'individu k et peut prendre dans notre étude deux modalités (0 ou 1). L'objectif de la discrimination est de prédire Y à partir des valeurs des X_i. Il faut pour cela construire une fonction

$\Phi : \begin{matrix} IR^p \to \{0,1\} \\ x \mapsto \Phi(x) = \hat{y} \end{matrix}$ appelée classificateur. Cette fonction attribue à un vecteur d'observations x une classe ŷ. Elle est déterminée à partir du jeu de données disponibles de telle sorte que $\Phi(x_k)$ soit le plus proche possible de y_k pour tout individu k. Dans notre étude, x_{ik} est la valeur du log ratio pour la séquence i et la souris k et y_k est la classe de la souris k, c'est-à-dire son traitement (ici rosiglitazone ou SCOMP). La discrimination consiste donc à déterminer avec quel produit une souris a été traitée à partir de l'observation de l'expression de ses gènes.

Lors de la construction d'un classificateur ϕ, certaines variables peuvent être redondantes ou inutiles pour la discrimination. Ceci est particulièrement vrai pour les données de puces à ADN qui sont caractérisées par un nombre très important de variables (plusieurs milliers de gènes) et peu d'observations (quelques dizaines d'échantillons). La sélection de variables consiste à choisir q variables parmi les p initiales de manière à construire le meilleur classificateur en termes de qualité de discrimination. Les méthodes de sélection de variables peuvent être regroupées en trois catégories : « filter », « wrapper » et « embedded » [34]. Les méthodes de type « filter » sélectionnent les variables intéressantes avant toute classification. Les méthodes de type « wrapper » utilisent la méthode de classification comme une boîte noire pour évaluer la qualité de sous-ensembles de variables. Enfin, pour les méthodes de type « embedded », la sélection de variables fait partie du processus d'apprentissage du classificateur. Dans notre étude, sélectionner les variables correspond à sélectionner les séquences participant le plus à la discrimination entre les deux traitements.

2.1.2 Evaluation de la qualité d'un modèle

Un modèle consiste en un choix de q variables parmi p et d'un classificateur ϕ construit à partir de ces q variables. La qualité d'un modèle est classiquement évaluée par sa qualité de discrimination entre les classes. On note Réel($\Phi r = P(\Phi(\Phi(\neq Y)$, le taux d'erreur ou risque réel associé à ϕ, où P désigne la loi de probabilité jointe du couple (X,Y). L'objectif est de minimiser ce risque réel. Pour cela, différents modèles sont testés et celui dont le taux d'erreur est minimal est conservé.

En pratique, la loi jointe du couple (X,Y) est inconnue et doit donc être estimée. L'estimateur par resubstitution du risque réel d'un classificateur ϕ est obtenu par le calcul du risque empirique : $Rresubs(\Phi) = Remp(\Phi) = \frac{1}{n}\sum_{k=1}^{n} I\{\Phi(x_k) \neq y_k\}$, où I_C est une fonction indicatrice valant 1 si la condition C est vraie et 0 sinon. L'estimateur par resubstitution est un estimateur simple mais biaisé du risque réel. En effet, en augmentant la complexité de ϕ, on peut artificiellement diminuer la valeur de l'estimateur par resubstitution sans pour autant garantir la qualité de la discrimination sur un nouveau jeu de données.

Plusieurs approches peuvent être envisagées pour obtenir un meilleur estimateur du risque réel. La méthode de hold-out consiste à séparer les données en deux échantillons : un échantillon d'apprentissage qui est utilisé pour construire ϕ et un échantillon de test qui est utilisé pour calculer le risque empirique. L'estimateur par hold-out est un estimateur non biaisé du risque réel mais il est dépendant du choix de l'ensemble d'apprentissage. Il est de plus difficile à utiliser quand peu d'observations sont disponibles. La validation croisée consiste à considérer plusieurs découpages des données en ensembles d'apprentissage et de test et à moyenner les erreurs par hold-out obtenues de manière à moins dépendre du choix de ces ensembles. Une autre technique est la construction de N classificateurs ϕ_j sur n individus tirés avec remise (échantillons bootstrap), c'est-à-dire que le nombre d'individus reste constant mais les individus peuvent être redondants. Chaque classificateur est testé sur les individus non sélectionnés dans l'échantillon bootstrap et l'estimateur du risque réel est donné par $\text{Rboot632} = \frac{1}{N}\sum_{j=1}^{N}(0.632 \times \text{Rboot}_j + 0.368 \times \text{Remp})$ où Rboot_j est l'erreur par hold-out de ϕ_j et Remp est le risque empirique du classificateur ϕ construit à partir de toutes les données. 0.368 correspond à la probabilité asymptotique qu'une observation n'ait pas été choisie dans un échantillon bootstrap.

Dans une comparaison de ces différentes méthodes d'estimation du risque réel d'un classificateur, Kohavi et al. préconisent l'utilisation d'une validation croisée avec un découpage en dix sous-ensembles [35]. Cette méthode n'est cependant pas adaptée dans notre étude du fait du faible nombre d'observation (12 souris) : la méthode de Leave-One-Out-Cross-Validation (LOOCV), plus adaptée à ce type de situation, a donc été utilisée. Une LOOCV consiste à construire un classificateur à partir de toutes les observations sauf une et à le tester sur l'observation restante. L'estimateur du risque réel est la moyenne des erreurs par hold-out calculées sur toutes les combinaisons possibles.

Lors de l'estimation du taux d'erreur d'un modèle où il y a eu sélection de variables, il est conseillé de découpler sélection de variable et calcul de l'erreur de manière à obtenir un estimateur non biaisé. Concrètement, il est recommandé de choisir les variables sur un ensemble d'apprentissage, puis de tester le classificateur obtenu sur un ensemble de test [36]. En pratique, quand l'estimateur par LOOCV est utilisé, cela suppose de sélectionner les variables pour chacun des n sous-groupes contenant n-1 individus. Si la méthode de sélection de variables est coûteuse en temps de calcul, cette approche est difficile à appliquer.

2.2 État de l'art des méthodes de discrimination et de sélection de variables appliquées aux puces à ADN

L'analyse de données de puces à ADN confronte les chercheurs à un problème de haute dimensionnalité. L'objectif est souvent de trouver un nombre restreint de gènes marqueurs d'une condition particulière (type de cancer, type de traitement). L'étude des différences entre rosiglitazone et SCOMP se rapporte au même problème, c'est-à-dire à la sélection de variables dans le cadre d'une discrimination entre deux conditions biologiques. Cette problématique a été abondamment traitée dans la littérature et les méthodes utilisées dans ce cadre sont développées dans ce chapitre.

2.2.1 Méthodes de discrimination

2.2.1.1 Classificateur de Bayes

On appelle classificateur de Bayes le classificateur ϕ^* tel que $\Phi^*(x_k) = \begin{cases} 1 \text{ si } P(Y=1/X=x_k) \geq P(Y=0/X=x_k) \\ 0 \text{ sinon} \end{cases}$, où P est la loi de probabilité jointe du couple (X,Y). Le classificateur de Bayes est théoriquement optimal au sens du risque réel : $\Phi^* = \underset{\Phi}{\text{Argmin}}(\text{Rréel}(\Phi))$. Néanmoins, en pratique la loi P n'est pas connue et $P(Y=1/X=x_k)$ doit être estimée. Le théorème de Bayes fournit la relation suivante : $P(Y=1/X=x_k) = \frac{P(Y=1)P(X=x_k/Y=1)}{P(Y=0)P(X=x_k/Y=0)+P(Y=1)P(X=x_k/Y=1)}$. Les différents paramètres sont estimés par maximum de vraisemblance à partir du jeu de données en supposant les variables X_i indépendantes.

Ce type de classificateur est rapide et peut traiter de larges bases de données. Le modèle probabiliste est simple, mais tant que la probabilité de la bonne classe est plus élevée que celle de l'autre classe, la classification reste exacte même si les probabilités n'ont pas été correctement estimées. Il demeure néanmoins difficile de traiter des situations où le nombre d'observations est largement inférieur au nombre de variables quand les variables dépendent les unes des autres. Le classificateur de Bayes a été comparé par Li et al. à d'autres méthodes de discrimination dans le cadre d'une application aux données de puces à ADN [37]. Cette étude a montré que le classificateur de Bayes peinait à classer correctement des tissus cancéreux. Excepté sur quelques jeux de données particuliers, ses performances étaient moindres par rapport à des méthodes de discrimination plus élaborées.

2.2.1.2 Analyse discriminante

L'analyse discriminante est fondée sur la formule de Bayes citée précédemment et suppose que X a une distribution gaussienne dans chacune des classes pour estimer $P(Y=1/X=x_k)$ et $P(Y=0/X=x_k)$. On se ramène alors à un problème d'estimation des paramètres des lois normales pour chaque classe : moyennes et matrices de covariance. Ces valeurs permettent de définir une frontière D entre les deux classes, combinaison des variables X_i initiales, qui définit le classificateur suivant : $\Phi(x_k) = \begin{cases} 1 \text{ si } D(x_k) \geq 0 \\ 0 \text{ sinon} \end{cases}$. Dans le cas général, cette frontière est quadratique et l'analyse est dite discriminante quadratique. Dans le cas particulier où les matrices de covariances sont égales dans les deux classes, on parle d'analyse discriminante linéaire ou d'analyse discriminante de Fisher et la frontière est de la forme : $B_0 + Bx = 0$. Si la matrice de covariance commune est de plus diagonale, l'analyse est discriminante linéaire diagonale.

Ce type de méthodes nécessite une adaptation quand il y a plus de variables que d'observations. En outre, si les classes n'ont pas la même dispersion ou si une classe comporte plusieurs sous-nuages, les résultats sont dégradés. Des analyses discriminantes ont déjà été appliquées dans le cadre de la classification de cellules cancéreuses à partir de données de puces à ADN. Elles étaient couplées à des méthodes de sélection de variables pour pallier le manque d'observations. Dans une comparaison de Dudoit et al. ***[38]***, l'analyse discriminante de Fisher a obtenu de moins bons résultats en termes de discrimination que d'autres méthodes de type K plus proches voisins, mais l'analyse discriminante linéaire diagonale restait compétitive.

2.2.1.3 Régression logistique

La régression logistique est fondée sur l'hypothèse : $\ln\left(\frac{P(Y=1/X=x_k)}{P(Y=0/X=x_k)}\right) = B_0 + Bx_k$, qui correspond à une frontière linéaire entre les deux classes. Cette propriété est notamment vérifiée dans le cadre de l'analyse discriminante linéaire quand la distribution des données est binormale, mais elle concerne également d'autres types de distributions. Après transformation, on obtient : $P(Y=1/X=x_k) = \frac{e^{B_0+Bx_k}}{1+e^{B_0+Bx_k}}$. Les paramètres sont estimés par maximum de vraisemblance et un nouvel individu est affecté à la classe ayant la probabilité la plus grande. La régression logistique fait partie des méthodes semi-paramétriques puisque seul le rapport des densités conditionnelles est décrit de manière paramétrique.

Le succès de la régression logistique repose principalement sur la possibilité d'interpréter les résultats obtenus. Pour deux individus k et l, l'odd-ratio est défini par :

$OR = \frac{P(Y=1/X=x_l)P(Y=0/X=x_k)}{P(Y=1/X=x_k)P(Y=0/X=x_l)} = e^{B(x_l - x_k)}$. Si les deux individus ne diffèrent que sur la variable i de sorte que $x_{il} = x_{ik} + 1$, on obtient $OR = e^{B_i}$. Si la classe 1 correspond à un risque (cancer par exemple), ce risque est multiplié par e^{B_i} quand la variable i est incrémentée d'une unité. Cette interprétation doit néanmoins être prise avec précaution car l'impact de l'augmentation d'une variable est considéré toutes choses égales par ailleurs, ce qui est rarement le cas en pratique. Par ailleurs, la difficulté de traitement des données de puces à ADN du fait de la faible quantité d'observations par rapport au nombre de variables demeure et des méthodes de sélection de variables doivent être mises en place en parallèle.

2.2.1.4 K plus proches voisins

La méthode des K plus proches voisins (K-Nearest Neighbours ou KNN) applique la règle suivante pour classer un nouvel individu k : les K plus proches voisins de k sont recherchés dans l'ensemble d'apprentissage, puis le nouvel individu reçoit la classe majoritaire parmi ces voisins (Figure 2.1). Cette approche consiste en fait à classer les individus en utilisant la quantité $P(Y=1/X \in V_K(x_k)) \approx P(Y=1/X=x_k)$, où $V_K(x_k)$ est le voisinage de l'individu k contenant K individus.

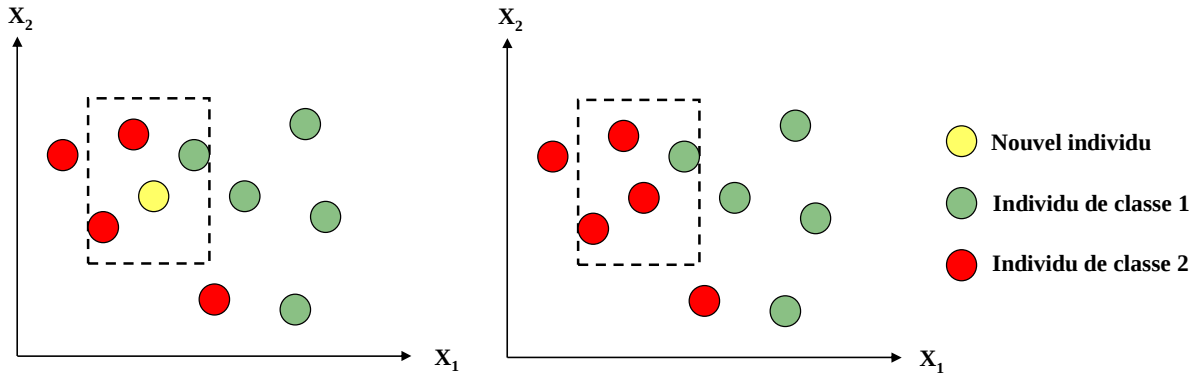

FIG. 2.1 : Principe des K plus proches voisins
A gauche : avant classification. A droite : après classification
Un nouvel individu est affecté à la classe de ses K plus proches voisins (ici K=3)

La méthode de KNN ne nécessite pas d'hypothèse sur la loi jointe de X et Y et ne rencontre pas de problème d'estimabilité quand il y a plus de variables que d'observations. Sa précision peut néanmoins être dégradée pour un trop grand nombre de variables : en grande dimension, les voisins sont loin et l'estimation des probabilités à posteriori n'est alors plus vraiment locale. Il est donc important de coupler cette méthode à une sélection des variables. Par ailleurs, les KNNs sont faciles à implémenter mais nécessitent un nombre important de calculs de distances quand la taille de l'ensemble d'apprentissage augmente. C'est pourquoi

des méthodes d'optimisation sont utilisées pour réduire le nombre de distances calculées. Ce problème concerne cependant rarement le traitement de données de puces à ADN car le nombre d'observations y est très limité.

Le paramètre K est un paramètre de complexité : la complexité du classificateur et le biais du risque empirique augmentent quand K diminue. Ce paramètre peut être optimisé par validation croisée. Dans les applications aux données de puces à ADN, caractérisées par un petit nombre d'observations, les valeurs de K sont classiquement faibles (1, 3 ou 5) et la distance euclidienne ou le coefficient de corrélation de Pearson sont utilisés (***[37][38][39][40][41][42]***). Dans une étude comparative de 2005 sur des méthodes de discrimination appliquées au traitement de données de puces à ADN pour la prédiction de cancers, les erreurs de classification des KNNs étaient sensiblement plus élevées que celles des méthodes de type Support Vector Machines détaillées plus loin ***[43]***. Après couplage avec des méthodes de sélection de variables, leurs résultats s'amélioraient. On peut noter que les KNNs ont obtenu des résultats à peu près équivalents à l'analyse discriminante linéaire diagonale ***[38]***.

2.2.1.5 Arbres de classification

Un arbre de classification est un modèle graphique dans lequel chaque nœud correspond à une variable (Figure 2.2). Un individu k est classé en fonction de ses réponses à des questions successives de type $x_{ik} > s_i$, où s_i est un seuil associé à la variable i. Les arbres de classification sont construits de manière récursive. La racine est déterminée par la recherche d'une question permettant de discriminer au mieux les deux classes à séparer. La qualité de discrimination correspond à la pureté en termes de classes des sous-ensembles d'observations générés par la réponse à cette question. La procédure est itérée pour construire les nœuds suivants jusqu'à ce que toutes les feuilles terminales soient pures. La classe d'un individu est déterminée par la feuille terminale qu'il atteint après avoir parcouru l'arbre de classification. Les arbres produits peuvent être binaires (ex : algorithme CART), mais peuvent également être plus larges. La complexité du classificateur augmente avec la taille de l'arbre, ce qui augmente également le biais de l'erreur empirique. Un compromis doit donc être trouvé entre pureté des feuilles terminales et biais de l'erreur. Pour cela, un élagage est réalisé. Il consiste à choisir un sous-arbre de l'arbre complet optimisant un critère qui dépend de la taille du sous-arbre et du risque empirique. Les arbres de classification peuvent notamment être utilisés pour la sélection de variables : seules les variables les plus discriminantes sont en théorie affectées à un nœud.

Il est possible d'agréger les arbres de classification pour produire des forêts aléatoires : un grand nombre d'arbres est généré, par exemple à partir de différents échantillons de la population initiale, et la classe finale attribuée à un individu provient d'un vote entre ces arbres. Une telle combinaison permet de stabiliser les résultats obtenus par rapport à un arbre utilisé seul. Les arbres de classification sont peu flexibles pour modéliser des distributions complexes de l'espace des variables, mais ils sont plus rapides que les réseaux de neurones

présentés par la suite. Leur principal atout est leur facilité de lecture et d'interprétation. Les arbres de classification et les forêts aléatoires ont déjà été utilisés pour classer des cellules cancéreuses à partir de données de puces à ADN. L'étude comparative de Li et al. en 2004 a conclu qu'ils obtenaient de moins bons résultats de classification que les Support Vector Machines *[37]*.

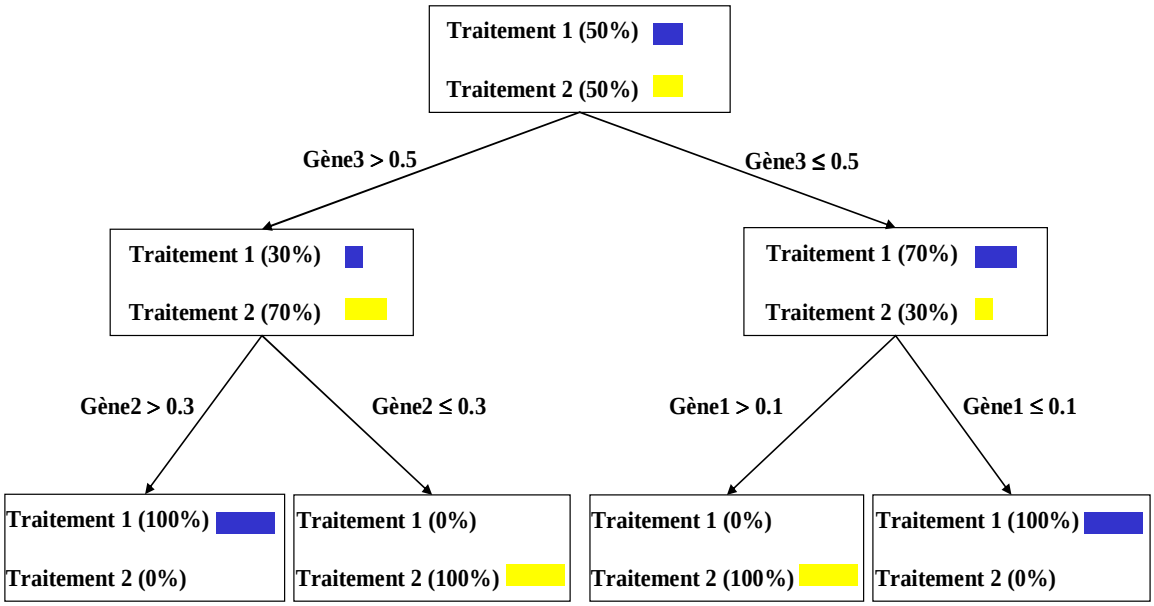

FIG. 2.2 : Exemple d'arbre de classification
La classe d'une observation est déterminée par la feuille terminale qu'elle atteint après avoir répondu aux questions successives représentées par l'arbre.

2.2.1.6 Réseaux de neurones artificiels

Un réseau de neurone artificiel est un classificateur qui se présente sous la forme d'un réseau de petites unités de calcul, les neurones. Les neurones sont reliés entre eux par des arêtes dirigées et pondérées. Chaque neurone calcule la somme pondérée de ses entrées, puis lui applique une fonction de transfert définie par l'utilisateur pour obtenir sa sortie (Figure 2.3). La fonction de transfert peut être une fonction seuil quand on recherche une variable qualitative ou bien une sigmoïde dans le cas de variables quantitatives. La structure du réseau est définie par l'utilisateur et détermine le type de fonction qu'il peut apprendre. Une forme classique des réseaux de neurones est le perceptron multicouche représenté sur la Figure 2.4. Les poids sont quant à eux optimisés par rétropropagation, c'est-à-dire qu'ils sont corrigés en fonction de l'erreur de sortie. Le modèle obtenu fournit la fonction Φ permettant de calculer la variable recherchée Y, à partir des variables X_i d'entrée.

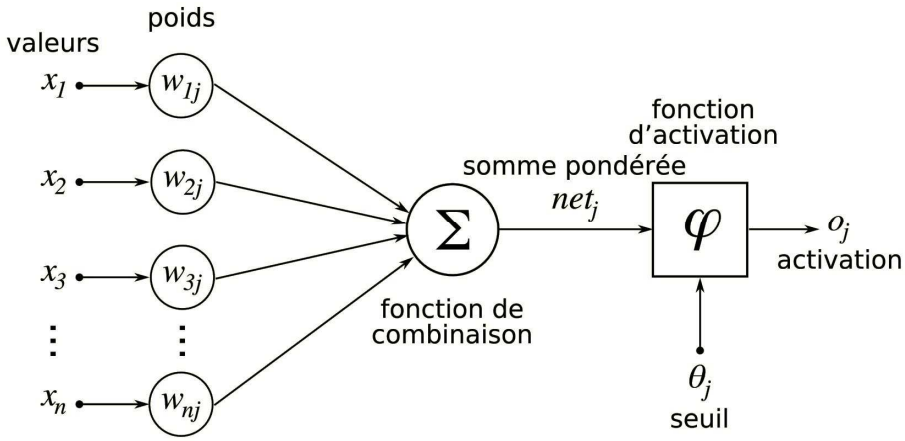

FIG. 2.3 : Unité de calcul [44]
Les valeurs d'entrée sont pondérées, puis sommées avant de passer par une fonction de transfert qui fournit la sortie du neurone

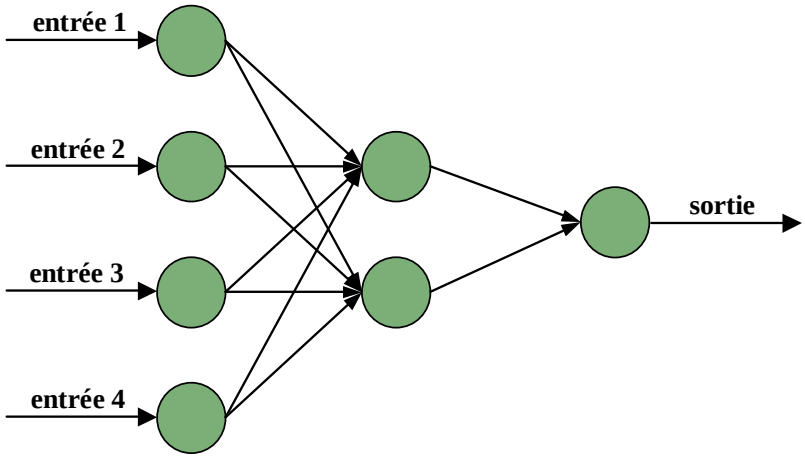

FIG. 2.4 : Perceptron multicouche

Les réseaux de neurones peuvent apprendre de nombreuses fonctions du fait de la diversité des architectures et des fonctions de transfert possibles. Cependant, l'architecture est difficile à optimiser. Il faut par ailleurs suffisamment de données pour pouvoir estimer les différents paramètres du modèle quand la structure est complexe (beaucoup de neurones ou beaucoup de liaisons). C'est pourquoi les réseaux de neurones considérés sont souvent simples et peuvent être couplés à des méthodes de sélection de variables. Les réseaux de neurones ont été utilisés dans l'identification et le contrôle de systèmes, la prise de décision dans des jeux, la reconnaissance de formes, etc.

Ils ont par ailleurs déjà été appliqués au traitement de données de puces à ADN dans le cadre de la prédiction de cancers. Dans la première application, la structure utilisée était très simple (pas de couche intermédiaire entre l'entrée et la sortie) et la méthode était couplée à une sélection des variables ***[45]***. La deuxième application est intégrée à une étude comparative ***[43]***. Les réseaux de neurones (une seule couche intermédiaire) ont réalisé des performances de classification globalement moins bonnes que des méthodes de type Support Vector Machines (SVM). Après couplage avec des méthodes de sélection de variables, il n'y avait pas de différence significative entre les différentes méthodes testées (réseaux de neurones, SVM, K plus proches voisins, …), même si les résultats obtenus par les SVMs étaient légèrement supérieurs.

2.2.1.7 Support Vector Machines ou Séparateurs à Vastes Marge

De même que pour l'analyse discriminante, le principe des Support Vector Machines (SVM) consiste à trouver une frontière D de l'espace des variables permettant de définir le classificateur suivant $\Phi(x_k) = \begin{cases} 1 \text{ si } D(x_k) \geq 0 \\ 0 \text{ sinon} \end{cases}$. Dans le cas des SVMs, cette frontière est construite de manière à minimiser l'erreur de classification obtenue sur un nouvel ensemble d'individus. L'idée s'illustre simplement sur un séparateur linéaire. Sur la Figure 2.5, les individus sont représentés dans le plan en fonction des deux variables X_1 et X_2. La classe Y est représentée par un symbole (cercle ou croix). Un SVM linéaire calcule l'équation d'un hyperplan (ici une droite) qui sépare les deux classes des données et maximise la marge entre lui-même et les individus les plus proches. L'équation de l'hyperplan est de la forme $D(x) = w.x + b$ où w est un vecteur de poids, b est une constante et x correspond aux coordonnées d'un individu dans l'espace des variables. Sur la Figure 2.5.a, toutes les droites en pointillés sont valides. Il est néanmoins intuitif que la droite en trait plein est optimale et plus susceptible de classer correctement un nouvel individu. La marge de sécurité introduite est portée par les observations les plus proches de l'hyperplan : ce sont les « vecteurs de support » (Figure 2.5.b) qui donnent leur nom à la méthode.

Les SVMs peuvent également être utilisés pour des cas de discrimination non linéaires. Cela nécessite le passage dans un espace de plus haute dimension de manière à se ramener à une séparation linéaire. On observe que, sur la Figure 2.6, passer de deux à trois dimensions permet de séparer les données de manière linéaire. La fonction φ qui réalise le changement de dimension est quant à elle non linéaire. Il n'est néanmoins pas nécessaire de la connaître pour définir un SVM. En effet, le calcul de l'hyperplan séparateur optimal nécessite seulement la connaissance du produit scalaire entre deux points images de φ. Il suffit donc de définir une fonction noyau à valeur réelle vérifiant $K(x_1, x_2) = \varphi(x_1).\varphi(x_2)$. Un SVM est alors uniquement caractérisé par sa fonction noyau. Les noyaux classiques sont linéaires, polynomiaux ou gaussiens.

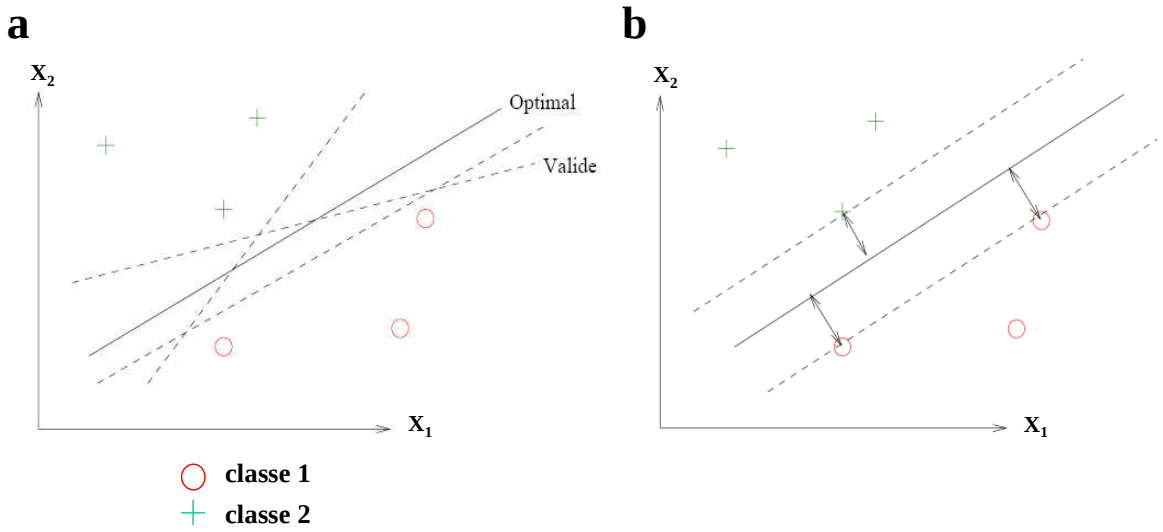

FIG. 2.5 : Principe des SVMs linéaires [46]
Un hyperplan optimal séparant les classes est déterminé en maximisant sa distance avec les individus les plus proches. Exemple en dimension 2.

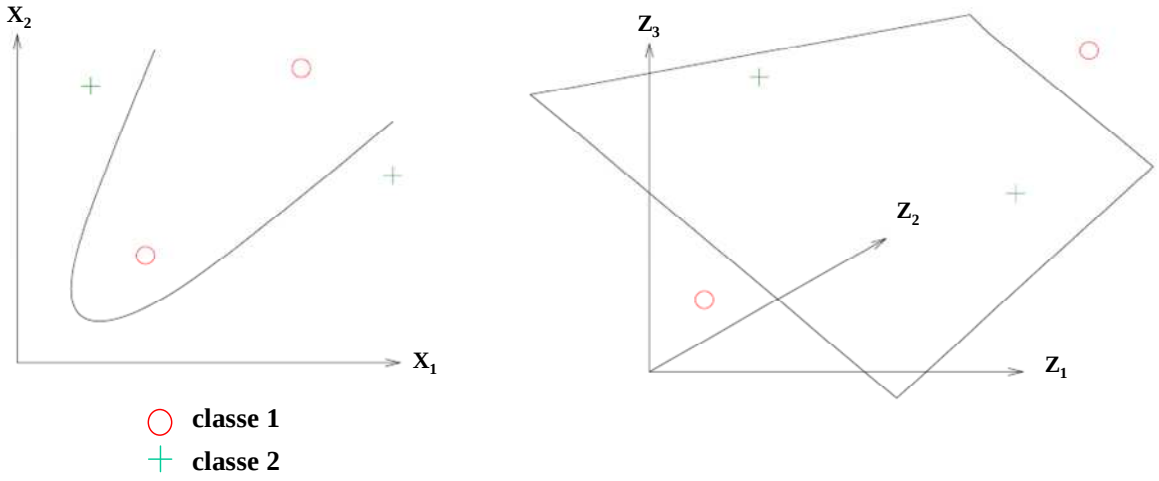

FIG. 2.6 : Exemple de SVM non linéaire [46]
Le passage dans un espace de dimension supérieure permet de se ramener à une séparation linéaire. Exemple d'un passage de dimension 2 en dimension 3.

Les SVMs ont été conçus pour discriminer deux classes mais il existe des variantes permettant de traiter le multicatégoriel. Un avantage important est l'absence de limite en terme de nombre de variables par rapport au nombre d'individus. Enfin, le sur-apprentissage est contrôlé par le système de « soft-margin » qui introduit une tolérance sur des échantillons mal classés lors de la détermination de l'hyperplan séparateur. Pour cela, la contrainte de séparabilité parfaite peut être relaxée, mais avec un coût choisi par l'utilisateur. Les SVMs ont une complexité polynomiale en le nombre de données d'apprentissage. De plus, la complexité de calcul du noyau est peu élevée par rapport à la dimensionnalité du problème. Les SVMs sont néanmoins sensibles au bruit introduit par des exemples mal classés à priori qui peuvent réduire considérablement leurs performances. Ils ont entre autres déjà été appliqués à la classification de textes, d'images et à la reconnaissance d'écriture manuelle. Il existe également des applications au traitement de données de puces à ADN dans le cadre de la discrimination entre types de cellules cancéreuses ([43][47]). Les noyaux utilisés sont essentiellement linéaires ou polynomiaux. Sans sélection de variables les taux d'erreur des SVMs étaient généralement inférieurs à ceux obtenus avec les autres méthodes. La sélection de variables réduit cet écart.

2.2.2 Méthodes de sélection de variables

2.2.2.1 Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode d'analyse descriptive qui est utilisée pour réduire la dimension de l'espace initial afin de visualiser les données en deux ou trois dimensions. Elle permet également de réduire l'espace des données avant l'utilisation d'une méthode de discrimination. Une nouvelle base de l'espace des données est construite. Ces nouvelles variables sont appelées composantes principales et sont calculées de manière à ce que chacune explique au mieux la variabilité restant dans les données. L'ACP n'est donc pas directement une méthode de sélection de variables, mais plutôt une méthode de réduction de l'espace.

Le principal inconvénient de l'analyse en composantes principales est la difficulté d'interprétation des résultats en fonction des variables initiales. L'ACP a déjà été utilisée dans le cadre de la sélection de variables pour la discrimination de cellules cancéreuses, mais la détermination des gènes était indirecte. Par exemple, en 2001, Khan et al. ont utilisé l'ACP uniquement comme méthode de réduction de l'espace et non comme méthode de sélection de variables [45]. Pour évaluer l'intérêt des gènes une étude de sensibilité a été réalisée : chaque gène était supprimé à tour de rôle, puis une ACP couplée à une méthode de discrimination était utilisée afin d'évaluer l'impact de ce retrait. Par ailleurs, une étude comparative de plusieurs méthodes de sélection de variables [48] a abouti à la conclusion que les résultats de la discrimination étaient meilleurs avec un T-test (détaillé ci-dessous) qu'avec une ACP.

2.2.2.2 Scores et tests statistiques

Il existe de nombreuses méthodes de classements unigènes, c'est-à-dire ne prenant pas en compte les interactions entre les variables, qui reposent sur des scores liés ou non à des tests statistiques. Dans les formules suivantes, x_{ik} est la réalisation de la variable i pour l'individu k, $\overline{x_{iq}}$ est la moyenne de la variable i sur les individus de la classe q et σ_{iq}^2 est sa variance. Enfin, $\overline{x_i}$ est la moyenne de la variable i sur tous les individus et n_q est le nombre d'individus dans la classe q. Les formules de deux principaux tests sont présentées ici dans le cas de deux classes (0 et 1).

- Le T-test avec correction de Welch permet de vérifier la significativité de la différence entre les moyennes de deux populations. Il est utilisé pour des échantillons de distributions normales et de variances différentes ou pour des échantillons de grande taille. Il est fondé sur le score T qui suit une loi de Student sous l'hypothèse d'égalité des moyennes :

$$T(i) = \frac{\overline{x_{i0}} - \overline{x_{i1}}}{\sqrt{\frac{\sigma_{i0}^2}{n_0} + \frac{\sigma_{i1}^2}{n_1}}}.$$

- Le F-test permet de vérifier si les moyennes de plusieurs populations sont égales. Les hypothèses sont les mêmes que pour le T-test. Il est fondé sur le score suivant :

$$F(i) = (n-2) \times \frac{\sum_{q=0}^{1} n_q (\overline{x_{iq}} - \overline{x_i})^2}{\sum_{q=0}^{1} n_q (n_q - 1) \sigma_{iq}^2}.$$

L'inconvénient principal de ce type de méthodes est l'absence de prise en compte des interactions entre les variables. Une variable peut avoir un mauvais score par elle-même mais être discriminante une fois combinée avec une autre variable. Par ailleurs, il ne faut pas oublier que toutes les hypothèses d'application des tests statistiques ne sont pas nécessairement respectées dans le cadre d'une application aux données de puces à ADN : normalité des distributions, égalité des variances. Ces méthodes sont donc souvent utilisées abusivement en pratique. On peut tout de même remarquer qu'une étude comparative [48] a présenté de meilleures performances de classification avec le T-test qu'avec d'autres scores unigènes.

2.2.2.3 Nearest Shrunken Centroids

La méthode de Nearest Shrunken Centroids (NSC) permet de sélectionner les variables dont la moyenne est différente dans une classe par rapport aux autres classes [49]. Les notations citées dans la partie précédentes sont reprises. $\overline{x_{iq}}$ est la moyenne de la variable i dans la classe q et est appelée centroïde. La moyenne de chaque variable, $\overline{x_i}$, est également calculée toutes classes confondues (centroïde global). En résumé, les valeurs $\overline{x_{iq}}$ sont tout d'abord normalisées par l'écart type intra-classe, puis elles sont contractées vers les centroïdes globaux $\overline{x_i}$.

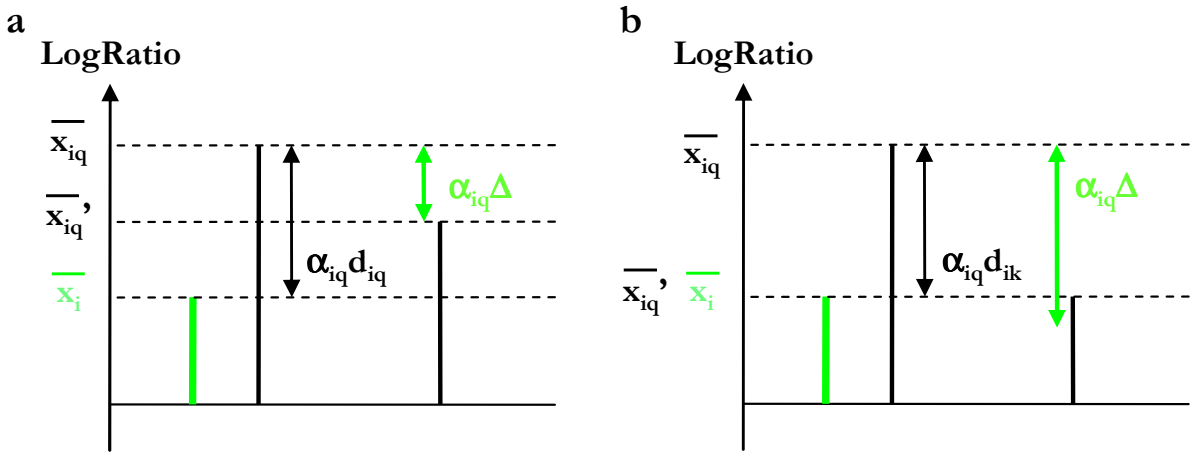

FIG. 2.7 : Principe de la méthode de Nearest Shrunken Centroids
(a) : Cas d'une variable discriminante. (b) : Cas d'une variable non discriminante.
Les centroïdes sont contractés vers les centroïdes globaux. Si la contraction est plus importante que l'écart entre le centroïde et le centroïde global, la variable est considérée comme non discriminante.

Soit $d_{iq} = \frac{\overline{x_{iq}} - \overline{x_i}}{\alpha_{iq}}$, où $\alpha_{iq} = \sqrt{\frac{1}{n_q} + \frac{1}{n}} \times (s_i + s_0)$ est un paramètre de normalisation. s_i est l'écart type intra-classes et s_0 est une constante égale à la médiane des s_i. La normalisation par les α_{iq} donne un poids plus élevé aux variables ayant une expression plus stable à l'intérieur d'une classe. On a alors $\overline{x_{iq}} = \overline{x_i} + \alpha_{iq} d_{iq}$, où d_{iq} est une t-statistique pour la variable i, comparant la classe q au centroïde global. Les nouveaux centroïdes contractés, $\overline{x_{iq}}'$, sont calculés pour chaque variable dans chaque classe : $\overline{x_{iq}}' = \overline{x_i} + \alpha_{iq} d_{iq}'$. d_{iq}' est obtenu en soustrayant à d_{iq} une quantité Δ en termes de valeur absolue et en conservant la partie positive

signée du résultat : $d_{iq}' = \text{sign}(d_{iq}) \times (|d_{iq}| - \Delta)_+$. Δ est le paramètre de la méthode qui permet de déterminer le nombre de variables conservées car les variables non discriminantes ont alors un d_{iq}' nul (Figure 2.7). Une nouvelle observation est affectée à la classe la plus proche en termes de centroïdes contractés. On peut noter qu'en fixant le paramètre Δ à zéro, on obtient simplement un classement des variables en fonction de leurs d_{iq} et que dans le cas de deux classes de même taille, les valeurs de d_{iq} obtenues pour les classes 0 et 1 sont opposées l'une de l'autre.

La méthode NSC présente le même inconvénient que les scores cités précédemment : elle néglige les interactions entre les variables. Le principe reste néanmoins intéressant puisque cette méthode a été développée dans le cadre de la discrimination entre types de cellules cancéreuses à partir de données de puces à ADN.

2.2.2.4 Significance Analysis of Microarrays

Significance Analysis of Microarrays (SAM) est une méthode développée pour détecter les gènes significativement régulés entre deux conditions biologiques [50]. Une différence relative d est calculée pour chaque variable i : $d_i = \frac{\overline{x_{i0}} - \overline{x_{i1}}}{s_i + s_0}$. Cette statistique est très proche de celle utilisée par NSC. Ici, s_i est une mesure de dispersion ($\sqrt{a\left(\sum_{k \in \text{classe0}} (x_{ik} - \overline{x_{i0}})^2 + \sum_{k \in \text{classe1}} (x_{ik} - \overline{x_{i1}})^2\right)}$, où $a = \frac{\frac{1}{n_0} + \frac{1}{n_1}}{n_0 + n_1 - 2}$) et s_0 est choisi de manière à minimiser le coefficient de variation de d. Les variables sont ensuite classées par valeurs de d croissantes : $d_{(1)} \leq d_{(2)} \leq ... \leq d_{(n)}$. En parallèle les données sont permutées b fois (permutation des observations), puis de nouvelles valeurs de d sont calculées pour chaque permutation et classées par ordre croissant. Pour chaque rang (i), la valeur moyenne de d est calculée ($\overline{d_{(i)}}$), fournissant des scores attendus dans le cadre d'une non-régulation. Les variables significativement régulées, sont les variables vérifiant $|d_{(i)} - \overline{d_{(i)}}| > \Delta$. Une estimation du taux de fausse découverte (espérance du taux de faux positifs) est ensuite fournie. Le paramètre Δ est choisi de manière à obtenir le taux de fausse découverte souhaité.

SAM est une méthode fréquemment utilisée en analyse de données des puces à ADN. Elle présente néanmoins deux limitations qui sont référencées dans une évaluation de SAM et d'un package de R associé [51]. D'une part, le taux de fausse découverte est surestimé dans le cas d'une trop grande dispersion des gènes non régulés. D'autre part, l'utilisation du seuil Δ peut mener à des résultats biaisés et même conflictuels quand les gènes régulées et non régulés ne sont pas bien séparés.

2.2.2.5 Independently Consistent Expression Discriminator

La méthode d'Independently Consistent Expression Discriminator (ICED) est à la fois une méthode de sélection de variables et de discrimination dans les cas où il n'y a que deux classes [52]. Elle recherche des variables consistantes, c'est-à-dire stable, dans une classe et non consistantes à la même valeur dans l'autre (Figure 2.8). Chaque variable i reçoit pour chaque classe q un poids W_{iq} (une variable a un poids élevé si elle est consistante dans une classe et non consistante à cette valeur dans l'autre). Une méthodologie de vote est ensuite mise en place pour la classification. Pour un nouvel individu k, un vote $V_q(x_{ik})$ est défini de la manière suivante pour chaque variable i : $V_q(x_{ik}) = W_{iq} \times \left| x_{ik} - \overline{x_{i\bar{q}}} \right|$, où $\bar{q}$ est la classe complémentaire de q et $\overline{x_{iq}}$ est la moyenne de la variable i sur la classe q. Une force de prédiction $P(x_k)$ est alors associée au nouvel individu à partir des ces votes pour p_0 meilleures variables de la classe 0 et p_1 meilleures variables de la classe 1 : $P(x_k) = \dfrac{p_1 \times \sum_{i=1}^{p_0} V_0(x_{ik}) - p_0 \times \sum_{i=1}^{p_1} V_1(x_{ik})}{p_1 \times \sum_{i=1}^{p_2} V_2(x_{ik}) + p_2 \times \sum_{i=1}^{p_1} V_1(x_{ik})}$. Si $P(x_k)$ est positif, l'individu k est affecté à la classe 0, sinon il est affecté à la classe 1. Une valeur absolue de P(x) élevée indique une bonne confiance dans la prédiction. p_0 et p_1 sont déterminés de manière à maximiser cette valeur absolue en moyenne sur tous les individus.

Les mêmes remarques que pour la méthode de Nearest Shrunken Centroids peuvent être faites. Le classement est unigène mais la méthode est originale puisqu'elle admet qu'un gène puisse être exprimé de façon aléatoire dans une situation donnée. Cette méthode a également été développée dans le cadre de la discrimination entre types de cellules cancéreuses à partir de données de puces à ADN.

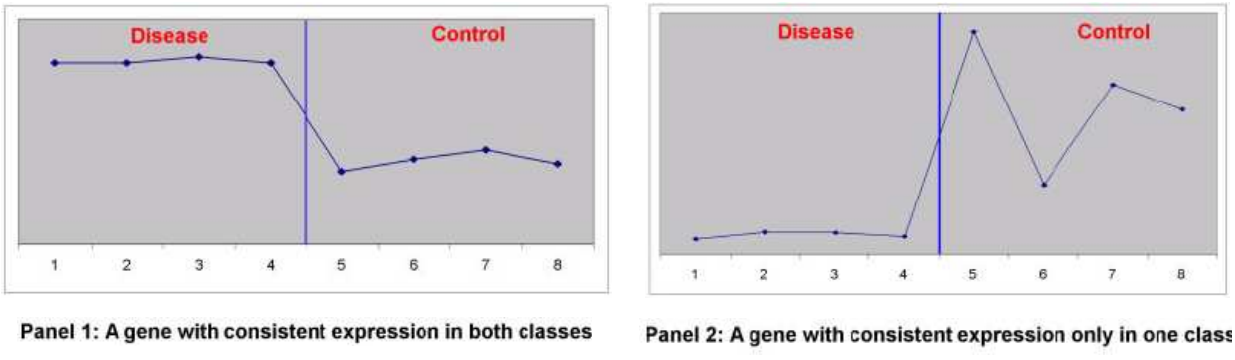

FIG. 2.8 : Comportement des variables sélectionnées par ICED (les variables sont des gènes) [52]
L'abscisse représente le numéro de l'individu et l'ordonnée le niveau d'expression du gène considéré.

2.2.2.6 Arbres de classification

Le principe des arbres de classification a déjà été expliqué précédemment. Les variables considérées comme discriminantes sont celles choisies pour servir de variables de segmentation aux nœuds de l'arbre. Les forêts aléatoires peuvent également fournir un ensemble de variables décisionnelles, par exemple en considérant les variables les plus représentées dans tous les arbres de classification construits. En 2004, une méthode de sélection de variables fondée sur une combinaison d'arbres de classification a été testée avec différents classificateurs dans une application à des données de puces à ADN [53]. Les taux d'erreur obtenus avec les gènes sélectionnés étaient meilleurs ou équivalents à ceux obtenus avec tous les gènes.

2.2.2.7 Recursive Feature Elimination

La Recursive Feature Elimination (RFE) correspond à l'élimination successive des variables apportant le moins à la qualité de la discrimination pour un classificateur donné. L'apprentissage est tout d'abord réalisé avec l'ensemble des p variables, la variable la moins discriminante est supprimée, puis l'apprentissage est réalisé sur les p-1 variables restantes et ce processus est itéré jusqu'à obtenir le nombre de variables désiré. Un cas particulier d'application de RFE est la Support Vector Machines – Recursive Feature Elimination (SVM-RFE). La méthode de discrimination utilisée est un SVM. L'équation du meilleur hyperplan séparateur est de la forme : $D(x) = w.x + b$ où w est un vecteur de poids, b est une constante et x correspond aux coordonnées d'un point dans l'espace des variables. Chaque variable X_i est associée à un poids (w_i^2) qui détermine son pouvoir discriminant à chaque itération.

La complexité de la RFE dépend directement du classificateur choisi et du nombre de variables éliminées à chaque itération. En terme d'application aux données de puces à ADN, les SVMs linéaires semblent être une méthode bien adaptée pour être combinée à la RFE ([47][54]). Par ailleurs, la SVM-RFE a déjà été combinée à une étude de fréquence afin d'évaluer l'importance des gènes sélectionnés [55] : des groupes de gènes prédicteurs ont été produits pour chaque ensemble d'apprentissage d'une validation croisée, puis une étude de la fréquence de sélection des gènes a été réalisée pour détecter les plus intéressants.

2.2.2.8 Algorithmes génétiques

Les algorithmes génétiques sont des méthodes d'optimisation. Ils sont utilisés dans le cadre de la sélection de variable pour optimiser le groupe des variables discriminantes sélectionnées. Ils permettent de parcourir de manière heuristique les sous-ensembles de variables possibles. Dans le cadre des algorithmes génétiques, un individu correspond à un sous-ensemble des variables initiales. Le principe des algorithmes génétiques est le suivant (Figure 2.9). Une population initiale d'individus est sélectionnée. Puis des hybridations et des mutations sont

effectuées au sein de cette population en fonction d'un score représentant la performance de discrimination des individus. Une hybridation correspond à un échange de variables entre individus et une mutation à l'insertion, la délétion ou la modification d'une variable au sein d'un individu. Enfin, un individu final, c'est-à-dire un sous-groupe de variables, est retenu quand un critère d'arrêt est atteint, par exemple un seuil d'erreur de classification.

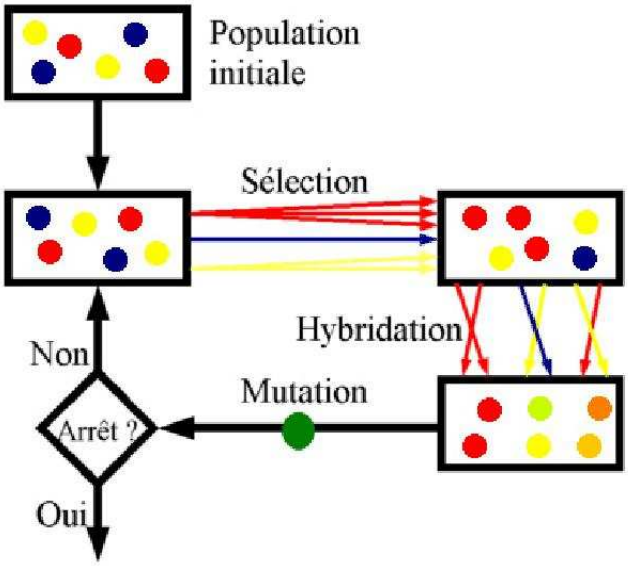

FIG. 2.9 : Principe des algorithmes génétiques [56]
Une population initiale de solutions subit hybridations et mutations en fonction de la qualité des solutions proposées. Après validation d'un critère d'arrêt, une meilleure solution est choisie.

Ce type d'algorithme permet de localiser une bonne solution qui peut néanmoins être un optimum local. Des règles très diverses peuvent être choisies pour les phases de sélection, d'hybridation et de mutation, ce qui peut rendre difficile le choix de l'algorithme ([40][41][42][57][58]). Par ailleurs, la complexité de l'algorithme croît rapidement avec la complexité de la fonction de score. Dans le cadre de la discrimination à partir de données de puces à ADN, la fonction de score correspond à la qualité de la discrimination par un classificateur donné. Typiquement, un algorithme génétique couplé à un séparateur bayésien est moins coûteux qu'un algorithme génétique couplé à un réseau de neurones artificiels. Globalement, les algorithmes génétiques ont souvent été couplés à la méthode des K plus proches voisins ou aux Support Vector Machines. En outre, ce type d'algorithme est par définition non déterministe et peut fournir plusieurs solutions différentes pour un même problème. Néanmoins, l'obtention possible de plusieurs groupes de gènes prédicteurs sous-optimaux peut permettre une étude de fréquence des gènes comme cela a été précédemment proposé pour la Support Vector Machines - Recursive Feature Elimination ([40][41]).

2.2.2.9 Couvertures de Markov

Y étant la variable à prédire, on appelle couverture de Markov de Y, notée MB(Y), l'ensemble minimal des variables X_i nécessaires à la prédiction de Y. Tout autre variable est conditionnellement indépendante de Y sachant MB(Y). La couverture de Markov correspond à ce que l'on recherche généralement dans le cadre d'une sélection de variables, mais elle est en pratique difficile à obtenir. En 2003, Aliferis et al. ont proposé un algorithme, HITON, capable de calculer une couverture de Markov sous réserve que certaines conditions soient vérifiées [59]. Le principe des couvertures de Markov est très intéressant, mais la réalisation des conditions nécessaires à leur calcul peut se révéler contraignante. En effet, la distribution jointe de X et Y doit être compatible avec un réseau bayésien et surtout l'ensemble d'apprentissage doit avoir une taille suffisante (supérieure à 150 d'après les auteurs), ce qui est rarement le cas pour les applications aux données de puces à ADN.

2.2.3 Choix de trois méthodes à tester

Au vu de cette étude bibliographique, cinq méthodes ont été présélectionnées pour être testées. Le T-test a été choisi, essentiellement comme référence, de part son usage largement répandu en statistiques. La méthode Nearest Shrunken Centroids (NSC) a été sélectionnée pour sa simplicité et son interprétabilité et la méthode Independently Consistent Expression Discriminator (ICED) l'a été pour son interprétabilité et sa conception dédiée aux puces à ADN. Les Support Vector Machines (SVM) permettent de gérer des espaces de variables importants et ont obtenu de bons résultats de classification sur des données de puces notamment en association avec la Recursive Feature Elimination (RFE). La SVM-RFE a donc également été sélectionnée. Par ailleurs, les K-plus proches voisins (KNN) ont été choisis pour leur simplicité et ont été associés à des algorithmes génétiques (GA) pour l'intérêt d'une approche d'optimisation : méthode appelée KNN-GA. Cependant, le programme de l'ICED s'est révélé indisponible et la méthode de KNN-GA a été éliminée en cours de test du fait de son caractère non déterministe et de son coût (détail en Annexe C avec les tests sur des lames de souris 22k). Au final, trois méthodes ont donc été sélectionnées pour être paramétrées, testées et potentiellement améliorées : T-test, NSC et SVM-RFE. Le T-test et la NSC sont des méthodes de type « filter », alors que la SVM-RFE est une méthode de type « embedded ».

2.3 Problématique mathématique

L'objectif principal de cette thèse est d'étudier les différences au niveau moléculaire entre deux agonistes PPAR, la rosiglitazone et le SCOMP. Il s'agit, au niveau mathématique, de trouver une méthode capable de sélectionner les gènes expliquant au mieux les différences entre les deux composés. Dans cette optique, trois méthodes ont été sélectionnées suite à une étude bibliographique : T-test avec correction de Welch, Nearest Shrunken Centroids et

Support Vector Machines - Recursive Feature Elimination. L'utilisation de ces méthodes comporte néanmoins certaines difficultés.

Un problème commun à ces méthodes est le manque de considération de l'impact de la variabilité technique des puces à ADN sur leurs résultats. Concrètement, on se demande dans quelle mesure les différences de résultats entre des puces à ADN réalisées à deux moments différents avec les mêmes échantillons influent sur les listes de séquences obtenues par les méthodes de sélection de variables. Dans une étude pour évaluer la reproductibilité inter et intra plate-formes des puces à ADN, le consortium MAQC a montré que leurs résultats étaient reproductibles pour des séquences présentant des rapports élevés entre intensité Cy5 et intensité Cy3 ou vice-versa (supérieurs à 1.4 ou 2) [60]. Cette information est très intéressante mais trop restrictive pour l'étude de gènes, certes peu régulés, mais tout de même impactés par un traitement. Un autre problème lié à ces méthodes est le choix du nombre de variables intéressantes. Par exemple, dans le cas de la méthode Nearest Shrunken Centroids, le nombre optimal de séquences est supposé être celui donnant la plus petite erreur de classification. Cependant, dans les cas où les classes sont bien séparées, l'erreur de classification est nulle même en considérant toutes les séquences.

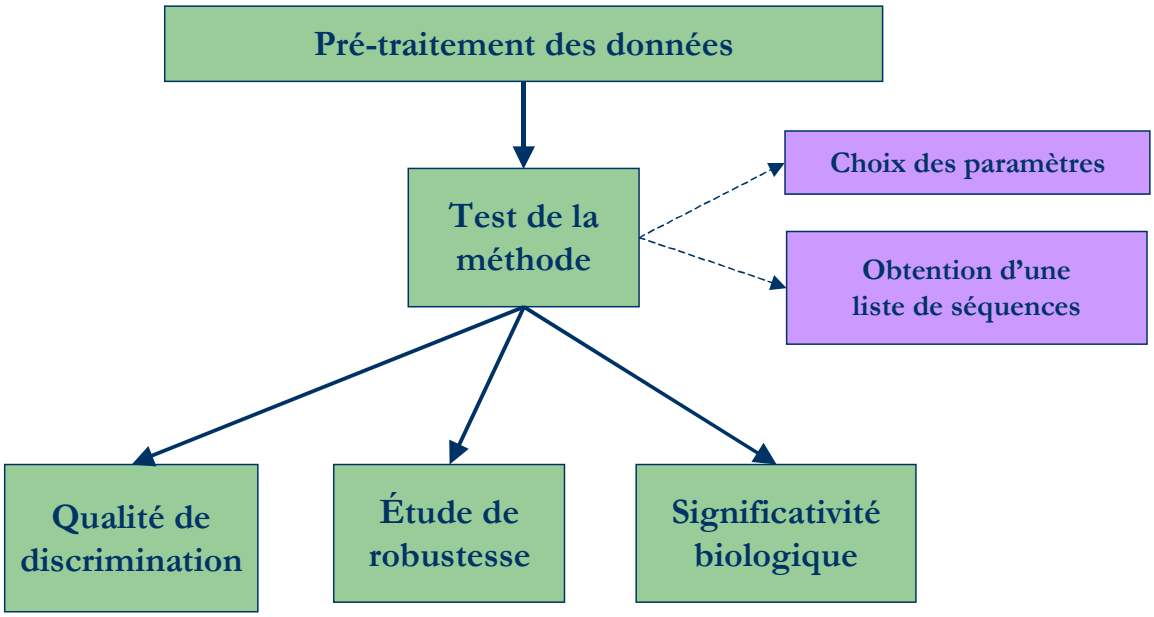

FIG. 2.10 : Protocole de test
Après pré-traitement des données, les méthodes de sélection de variables sont paramétrées et testées. Elles sont évaluées par la qualité de leur discrimination, leur robustesse par rapport à la variabilité technique des puces à ADN, et la significativité biologique de leurs résultats.

Plusieurs points doivent donc être considérés dans le cadre de la problématique mathématique :

- Quelle est la méthode la plus efficace parmi les trois sélectionnées ?
- Comment choisir un nombre de séquences optimal ?
- Comment tester et améliorer la robustesse par rapport à la variabilité technique des puces à ADN ?

Le protocole de test utilisé afin de tester les trois méthodes sélectionnées est présenté sur la Figure 2.10. Les données sont pré-traitées, puis les méthodes sont paramétrées et testées. La qualité des méthodes est évaluée par la qualité de la discrimination, la robustesse face à la variabilité technique des données de puces à ADN et la significativité biologique des listes de séquences obtenues. En parallèle, une nouvelle méthodologie a été développée afin de répondre aux questions de robustesse et de choix d'un nombre de gènes. Cette méthodologie est présentée dans le chapitre suivant.

Chapitre 3 :

Méthodologie robuste de sélection de gènes, MetRob

L'objectif principal de cette thèse est l'étude des différences au niveau moléculaire entre deux agonistes PPAR : la rosiglitazone et le SCOMP. Les deux composés ont été testés à trois doses, 28 µmol/kg, 84 µmol/kg et 280 µmol/kg, chez des souris diabétiques db/db. Trois organes, le muscle squelettique, le foie et le tissu adipeux inguinal, ont été prélevés et analysés en puces à ADN versus le groupe témoin de souris non traitées. Les agonistes PPAR ont été comparés par dose et par tissu conduisant à neuf jeux de données différents.

Au niveau mathématique, cette problématique correspond à trouver une méthode capable de sélectionner les gènes expliquant au mieux les différences entre les deux composés. Pour cela, trois méthodes de sélection de variables ont été choisies pour être testées : T-test, Nearest Shrunken Centroids et Support Vector Machine Recursive Feature Elimination (cf 2.2.3). Ces méthodes présentent néanmoins deux inconvénients majeurs qu'il convient de gérer : méconnaissance de l'impact de la variabilité technique des puces à ADN sur leurs résultats et difficulté à choisir le nombre de séquences à conserver. Une nouvelle méthodologie, MetRob, permettant la génération d'une liste de séquences robuste et reproductible et applicable à ces trois méthodes, a été développée afin de pallier ces problèmes. Une publication est en cours de soumission sur ces travaux [61].

Le principe de MetRob est présenté dans la première partie de ce chapitre. Les deux parties suivantes justifient et détaillent les choix de méthodologie et de paramètres qui ont été faits.

3.1 Principe global de la méthodologie MetRob

3.1.1 Pré-traitement des données

Les données ont nécessité une étape de pré-traitement avant de pouvoir appliquer les méthodes de sélection de variables. Les 41174 séquences représentées sur une puce à ADN ont tout d'abord été pré-filtrées afin d'éliminer celles qui n'étaient pas influencées par les

traitements. Pour chaque jeu de données, correspondant à une dose et un tissu, les séquences statistiquement significativement régulées (séquences SSR) ont été définies de la manière suivante. Une séquence est dite SSR si son expression pour un des deux traitements est statistiquement différente de l'expression du groupe témoin. Pour la comparaison de la rosiglitazone et du SCOMP, une p-value (définie au 1.5.1) inférieure ou égale à 0.01 pour au moins la moitié des souris d'un groupe traité a été considérée comme significative (Figure 3.1). Ce seuil de p-value est couramment utilisé pour les analyses de données issues de la plate-forme de puces à ADN. Il fournit des listes de séquences validées par d'autres technologies comme la qPCR (quantitative Polymerase Chain Reaction) qui permet une quantification absolue de la quantité des ARNs cibles. Aucun filtre sur les log ratios n'a été appliqué afin de ne pas exclure des séquences certes peu régulées, mais potentiellement discriminantes.

Les valeurs manquantes rapportées par le logiciel Feature Extraction (spots saturés ou spots ayant un signal non uniforme) ont été traitées en utilisant une méthode des K plus proches voisins. L'objectif était de permettre aux algorithmes de fonctionner tout en influençant le moins possible les résultats. Chaque valeur manquante a été remplacée par la moyenne de la variable correspondante dans les K plus proches observations (distance euclidienne). K a été choisi égal à 3, c'est-à-dire à la moitié des souris pour un traitement. Les données ont de plus été classiquement centrées et normées de manière à ce que le log ratio de chaque séquence ait une moyenne nulle et une variance unité.

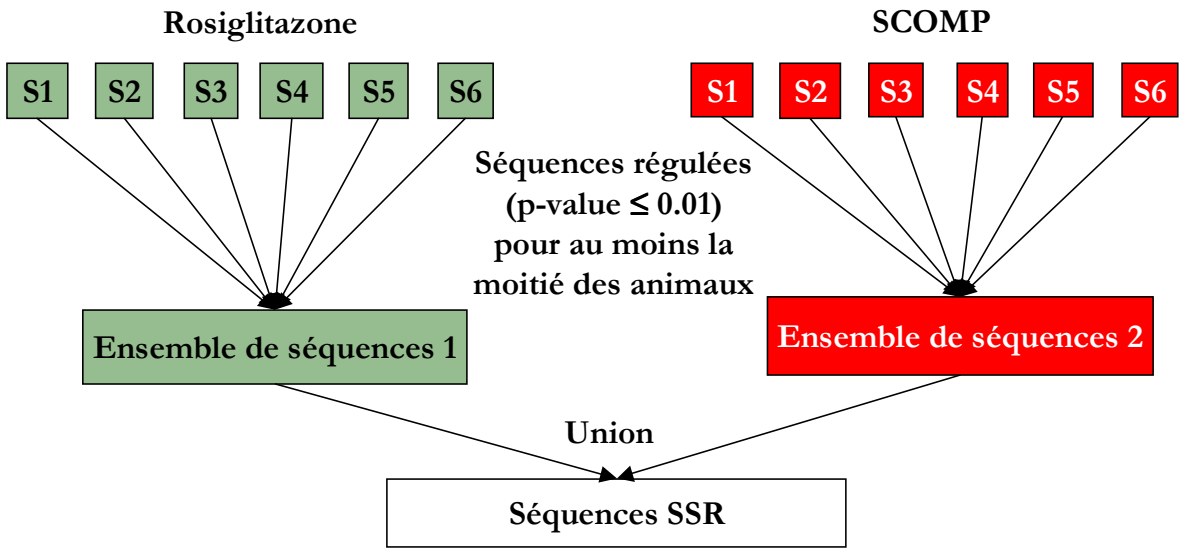

FIG. 3.1 : Définition des séquences statistiquement significativement régulées (séquences SSR)

3.1.2 Définition de la robustesse

MetRob a été conçue afin de prendre en compte l'impact de la variabilité technique des puces à ADN sur les résultats des méthodes de sélection de variables. Pour cela, nous avons choisi d'introduire la notion de robustesse d'une liste de séquences. Cette robustesse a été évaluée comme la stabilité de la liste de séquences obtenue avec une méthode sous l'effet de perturbations des données d'entrée. Les trois méthodes ont été comparées en termes de robustesse des listes de séquences qu'elles génèrent lorsque les perturbations considérées sont liées à la variabilité technique des puces à ADN. Comme il n'est économiquement pas possible de réitérer les expériences un grand nombre de fois, la comparaison a requis de générer des jeux de données virtuels reproduisant cette variabilité technique.

La robustesse d'une liste de séquences a été définie de la manière suivante. Une liste de L séquences est obtenue à partir des données initiales non perturbées. K listes de L séquences sont obtenues à partie de K jeux de données perturbées. Chaque liste sur données perturbées est comparée à la liste de séquences initiale en termes de pourcentage de séquences communes. La moyenne de ce pourcentage sur les K listes obtenues à partir de données perturbées est appelée robustesse. Cette robustesse a été utilisée à la fois pour évaluer la qualité des méthodes de sélection de variables et pour choisir un nombre de séquences à conserver optimal.

3.1.3 MetRob

La méthodologie MetRob permet de générer une liste de séquences robuste et reproductible expliquant au mieux les différences entre deux traitements. Elle peut être utilisée avec chacune des trois méthodes de sélection de variables à tester : T-test, Nearest Shrunken Centroids (NSC) et Support Vector Machine – Recursive Feature Elimination (SVM-RFE). Son principe est illustré sur la Figure 3.2.

Les données sont pré-traitées et les valeurs des log ratios sont considérées. La méthode de sélection de variables est utilisée pour classer toutes les séquences à partir du jeu de données initial. Cette méthode est ensuite utilisée pour classer les séquences à partir de 300 jeux de données perturbées. La technique de perturbation des données a été choisie en accord avec une étude de la variabilité technique des puces à ADN, de manière à obtenir des résultats cohérents avec la réalité (voir partie 3.2 pour l'étude de variabilité technique et les tests de perturbations). La robustesse est ensuite calculée, comme défini ci-dessus, pour chaque longueur de liste de séquences avec un pas de 10 séquences : la robustesse est calculée pour les 10 meilleures séquences, puis pour les 20 meilleures séquences, etc. Une longueur de liste de séquences est choisie de manière à maximiser la robustesse tout en gardant un nombre de séquences minimal (méthode détaillée dans la partie 3.3.4). On obtient donc une liste Λ de séquences sur données non perturbées et 300 listes de séquences sur données perturbées. La

liste de séquences finale est définie comme les séquences de Λ présentes dans au moins 80% des listes obtenues à partir des données perturbées (voir partie 3.3.6).

Python 2.4.4 a été utilisé pour implémenter MetRob (pré-traitement des données, perturbation des jeux de données…). Le logiciel R (version 2.4.1) a permis d'effectuer le T-test ainsi que la méthode NSC via le package PAM [62]. La toolbox Spider de Matlab (version R2007a) a été utilisée pour la SVM-RFE [63]. Enfin, la connexion entre les différents programmes a été réalisée via Python et les modules rpy [64] et mlabwrap [65].

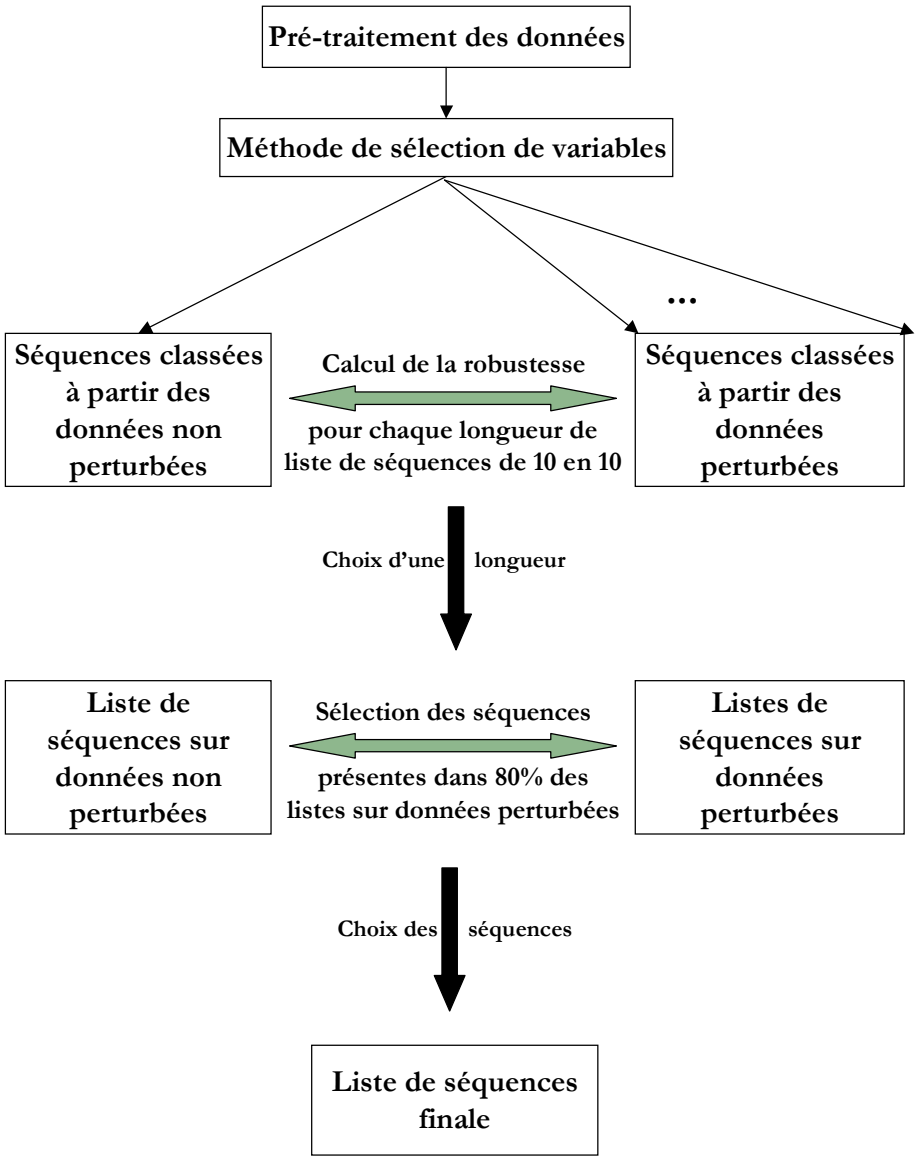

FIG. 3.2 : Principe de MetRob

3.2 Perturbation des données

MetRob nécessite de perturber artificiellement les données afin de simuler la variabilité technique des puces à ADN et d'évaluer la robustesse des méthodes de sélection de variables. Une méthode de perturbation des log ratios cohérente avec la réalité a donc été recherchée.

3.2.1 État de l'art

Différentes formes de bruit ont déjà été appliquées aux puces à ADN dans la littérature. En 2002, McShane et al. ont utilisé un bruit blanc gaussien pour perturber les valeurs des log ratios. La variance était estimée à partir des données en utilisant un centile de la distribution des variances [66]. Par la suite, Sayyed-Ahmad et al. ont proposé un autre type de perturbation. Ils ont ajouté au log ratio un bruit BLR_i de la forme suivante pour toute séquence i : $BLR_i = \text{LogRatio(i)} \times (2r-1) \times c$, où r est tiré au hasard entre 0 et 1 et où 100×c est le coefficient de variation des données [67] (Figure 3.3).

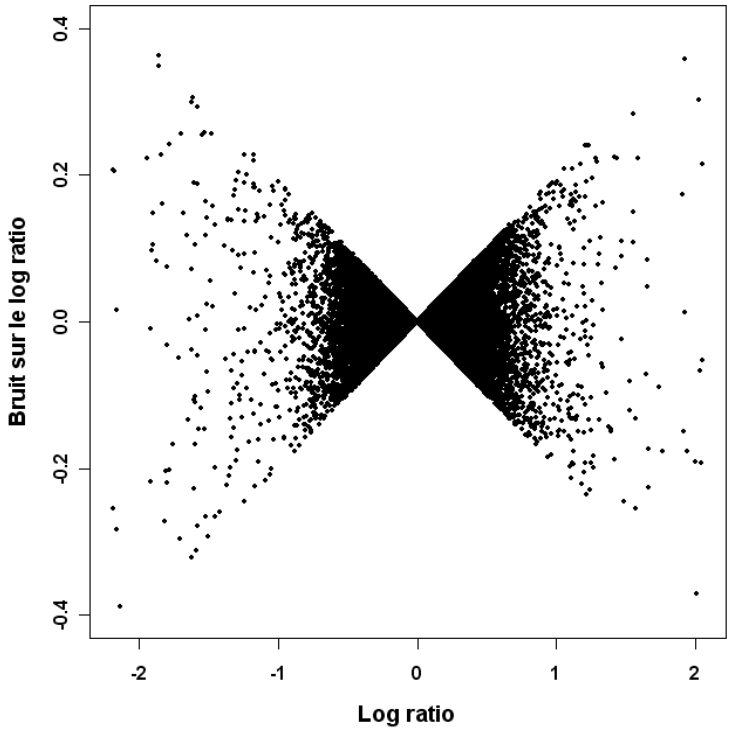

FIG. 3.3 : Allure de la perturbation proposée par Sayyed-Ahmad et al. pour un coefficient de variation des données de 20%
Bruit sur le log ratio en fonction du log ratio. La perturbation a été appliquée aux log ratios des lames de l'étude de variabilité technique (cf 1.6.2)

Estimer la variance du bruit directement à partir des données expérimentales de l'étude conduit à considérer la variabilité biologique. Ces deux approches nécessitent donc un jeu de données de réplicats techniques pour évaluer la variabilité technique qui est recherchée. Une étude de reproductibilité des puces à ADN à grande échelle a été menée par le consortium MAQC en 2006 ***[68]***, mais elle ne comportait pas les lames de souris Agilent 4*44k qui ont été utilisées dans le cadre de cette thèse. Une étude de variabilité technique a donc été réalisée en interne. Nous verrons par la suite que les données de variabilité que nous avons obtenues ne vérifient pas les deux types de relation proposés.

3.2.2 Etude de variabilité technique

L'étude de variabilité technique effectuée en interne sur les lames de souris Agilent 4*44k est développée dans le Chapitre 1 (cf 1.6.2). Deux échantillons biologiques ont été utilisés et six réplicats techniques ont été réalisés par échantillon. Pour cette étude, seules les séquences ayant une p-value inférieure ou égale à 0.05 pour au moins un réplicat technique ont été sélectionnées. Le seuil est moins strict que celui de la comparaison entre rosiglitazone et SCOMP afin de ne pas être trop restrictif sur le bruit. Le même type d'étude avait été mis en place avec les anciennes lames de souris 22k utilisées pour des expériences sur des souris ob/ob (Annexe A). Les résultats de cette étude sont fournis en Annexe C.

Pour chaque séquence i et chaque réplicat technique j, le bruit sur le log ratio BLR_{ij} est défini comme la différence entre le log ratio de l'observation considérée j et la moyenne des log ratios sur tous les réplicats techniques. Les bruits sur les intensités Cy3 (verte) et Cy5 (rouge), BIV_{ij} et BIR_{ij}, sont définis de manière similaire. Afin de reproduire cette variabilité pour perturber les données de puces à ADN, il faut tout d'abord en étudier les propriétés.

3.2.2.1 Bruit sur le log ratio

La distribution du bruit sur le log ratio a tout d'abord été considérée (Figure 3.4). Il ne s'agit visiblement pas d'une distribution normale, ce qui est confirmé par un test de Shapiro-Wilk pour lequel des p-values significatives inférieures à 0.01 ont été obtenues avec plusieurs échantillons de taille 50 (détail du test en Annexe D). L'hypothèse de normalité est donc rejetée. Par conséquent, la perturbation proposée par McShane et al. ne peut pas s'appliquer dans notre étude. Ce type de distribution semble néanmoins proche d'une loi normale au carré signée, loi qui a donc fait partie des tests de perturbations présentés ultérieurement.

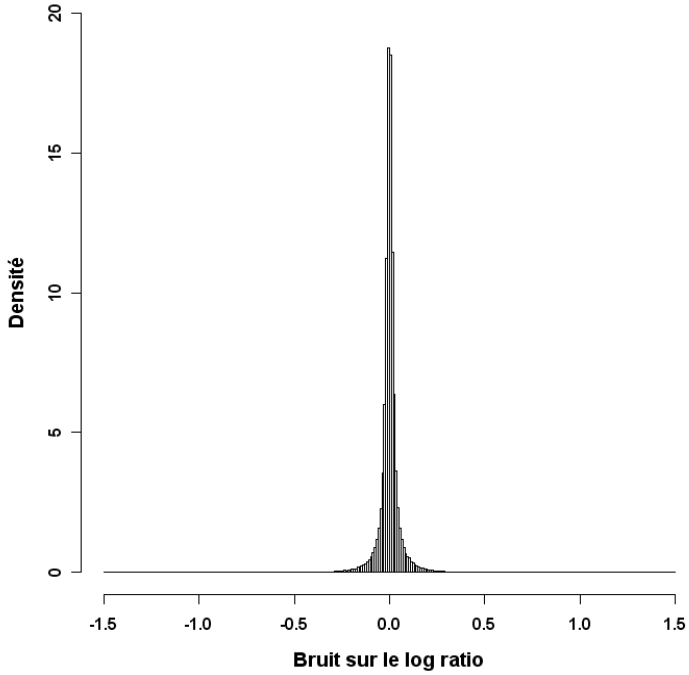

FIG. 3.4 : Histogramme du bruit sur le log ratio
Les barres sont de largeur 0.01. Les deux échantillons biologiques sont regroupés.

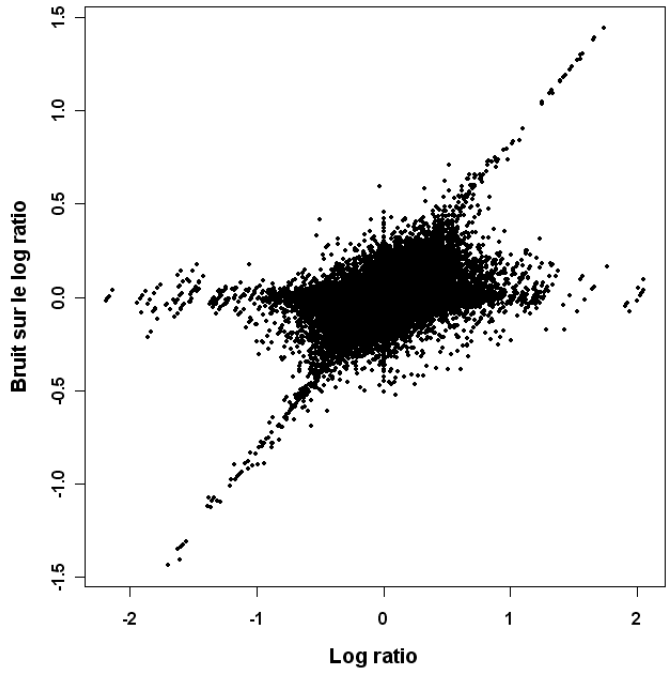

FIG. 3.5 : Bruit sur le log ratio en fonction du log ratio
Les deux échantillons biologiques sont regroupés.

Dans un second temps, le bruit sur le log ratio a été étudié un peu plus précisément, en recherchant une dépendance par rapport à la valeur du log ratio. Il ne semble cependant pas y avoir de relation exploitable entre le bruit sur le log ratio et le log ratio (Figure 3.5). Seuls quelques points montrent une relation linéaire entre les deux valeurs. Ces points correspondent aux séquences i ayant une valeur de log ratio nulle (cf explications dans l'Annexe B, paragraphe B.1.7) pour tous les réplicats techniques sauf le réplicat j. Dans ce cas, la moyenne M_i des log ratios pour les six réplicats techniques est donnée par $M_i = \frac{\text{LogRatio}_j}{6}$. On obtient donc $BLR_{ij} = \text{LogRatio}_j - M_i = \frac{5}{6}\text{LogRatio}_j$, ce qui explique la relation de linéarité. Il est par conséquent assez clair que la perturbation proposée par Sayyed-Ahmad et al. n'est pas adaptée à notre problématique en ce qui concerne le bruit sur le log ratio.

3.2.2.2 Lien avec les intensités

Les valeurs mesurées directement sur les puces à ADN sont les intensités. Nous avons donc envisagé de perturber les intensités Cy3 et Cy5, puis de recalculer un log ratio à partir de ces valeurs perturbées. Par conséquent, le bruit sur les intensités a également été étudié.

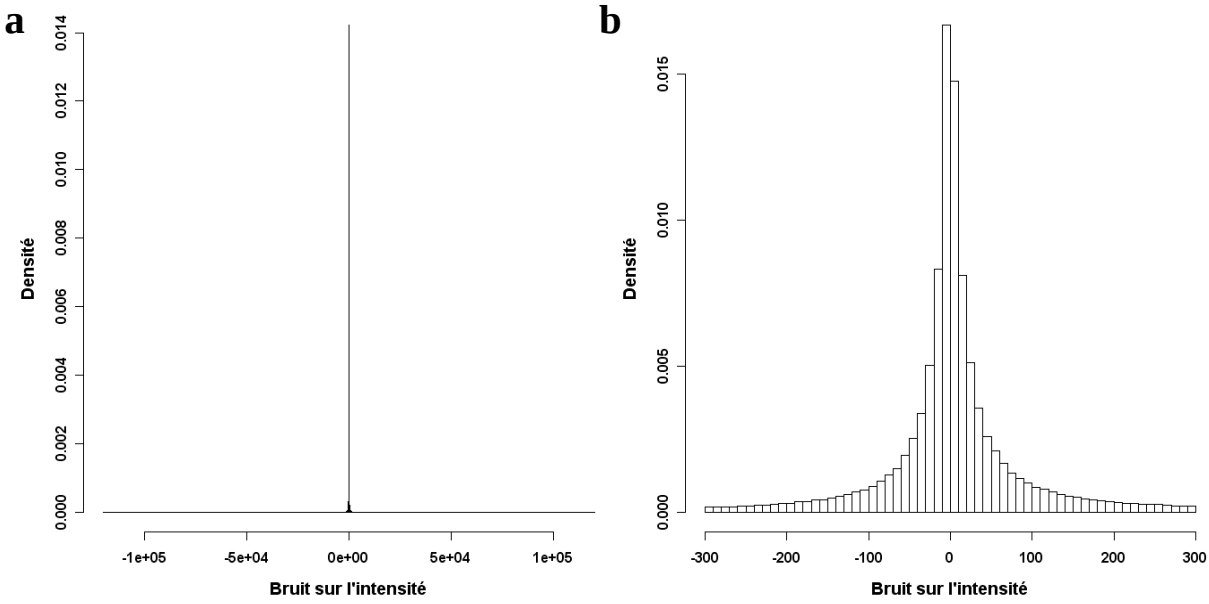

FIG. 3.6 : Histogrammes du bruit sur l'intensité
(a) : Histogramme avec toutes les valeurs d'intensité. (b) : Histogramme avec les valeurs d'intensités comprises entre –300 et 300. Pour les deux histogrammes, les barres sont de largeur 10. Les deux échantillons biologiques et les deux intensités (Cy3 et Cy5) sont regroupés.

Comme pour le bruit sur le log ratio, on observe sur la Figure 3.6.a que le bruit sur l'intensité ne suit pas une loi normale (p-values inférieures à 0.01 avec le test de Shapiro-Wilk sur des échantillons de taille 50). Le deuxième histogramme du bruit sur les intensités est présenté pour des valeurs comprises entre –300 et 300, représentant environ 85% des valeurs initiales, afin de mieux discerner la forme de la distribution (Figure 3.6.b). La relation entre bruit sur l'intensité et valeur de l'intensité a également été étudiée (Figure 3.7.a). L'écart type du bruit augmente avec l'intensité et le graphe ressemble cette fois beaucoup plus au côté positif de la perturbation proposée par Sayyed-Ahmad. Enfin, il faut noter que les bruits sur les intensités Cy3 et Cy5 semblent être liés linéairement (Figure 3.7.b).

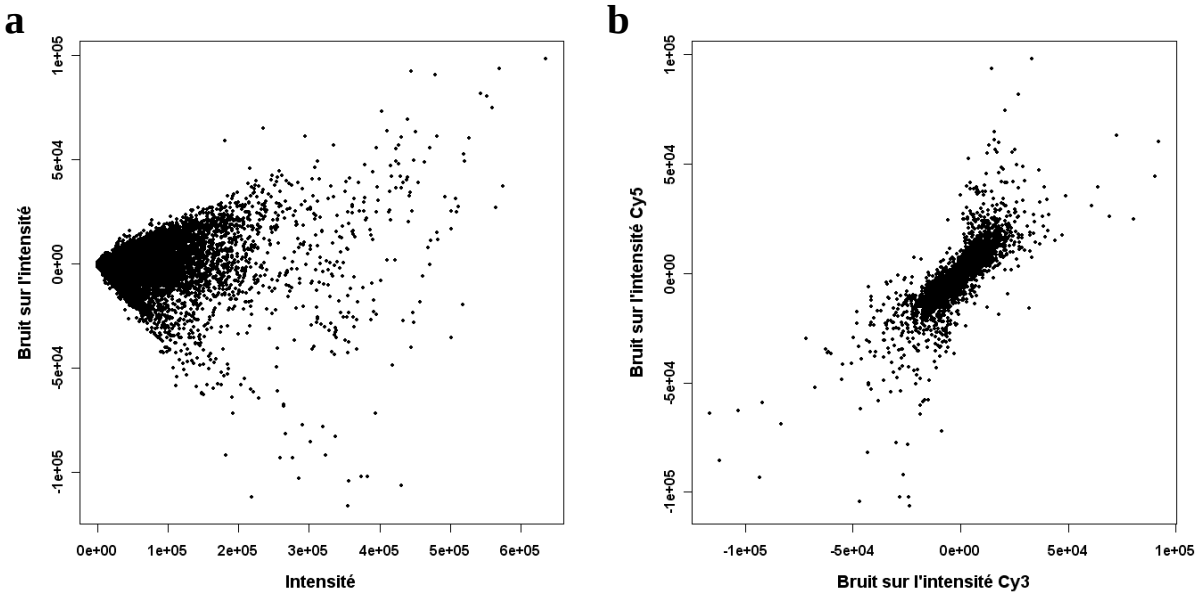

FIG. 3.7 : Propriétés du bruit sur l'intensité
(a) : Bruit sur l'intensité en fonction de l'intensité. Les deux échantillons biologiques et les deux intensités (Cy3 et Cy5) sont regroupés. (b) : Bruit sur l'intensité Cy5 en fonction du bruit sur l'intensité Cy3.

D'autre part, puisque le bruit sur le log ratio est recherché, mais que la mesure effectuée directement sur la lame concerne les intensités, une dernière approche a consisté à étudier la relation entre bruit sur le log ratio et intensités (Figure 3.8). On observe sur la Figure 3.8.a que le bruit est beaucoup plus variable pour de faibles intensités. La Figure 3.8.b, qui permet d'observer les intensités supérieures à 125 à une échelle plus précise, montre que l'écart type du bruit diminue dans un premier temps avec l'intensité, avant de réaugmenter très légèrement sur des hautes intensités. On peut noter que le bruit sur les hautes intensités est beaucoup plus faible avec les nouveaux protocoles des lames 4*44k qu'avec les anciennes lames 22k (voir Annexe C).

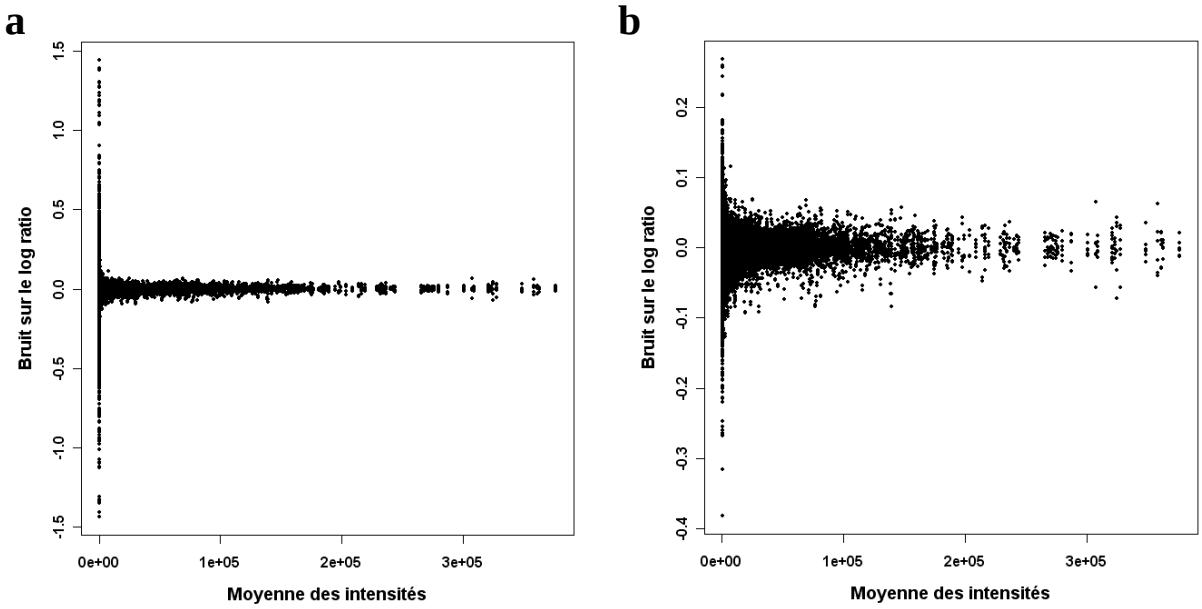

FIG. 3.8 : Bruit sur le log ratio en fonction de la moyenne des intensités Cy3 et Cy5
(a) : Toutes les valeurs d'intensité. (b) : Valeurs d'intensité supérieures à 125. Pour les deux graphes, les valeurs d'intensité considérées ont également été moyennées sur les réplicats techniques et les deux échantillons biologiques sont regroupés.

Toutes ces observations ont conduit au test de différentes formes de perturbations, afin de reproduire le bruit lié à la variabilité technique des puces à ADN. Les résultats de ces tests ainsi que le choix de la perturbation finalement utilisée sont présentés dans la partie suivante.

3.2.3 Test de différentes perturbations

3.2.3.1 Perturbation 1 : perturber directement le log ratio

La première perturbation testée repose sur l'observation de la distribution du bruit sur le log ratio dans l'étude de variabilité technique (Figure 3.4). Cette distribution n'est pas normale mais pourrait ressembler à une loi normale au carré signée. Les données ont donc été perturbées en ajoutant une variable aléatoire $V = \text{Signe}(U) \times U^2$, où U suit une loi normale centrée en zéro. L'écart type de U a été calculé à partir des données de l'étude de variabilité : $\sigma_U = \sqrt{\frac{\sigma_{\text{EtudeVar}}}{\sqrt{3}}} = 0.18$, avec σ_{EtudeVar} l'écart type du bruit observé sur les données de variabilité technique (voir le calcul de la formule en Annexe D).

La distribution de bruit obtenue est représentée sur la Figure 3.9. Elle ressemble effectivement à celle observée pour le bruit technique réel avec un étalement cependant moins large. Un test de Kolmogorov-Smirnov conclut que l'hypothèse nulle de même distribution initiale pour le bruit réel et le bruit obtenu avec la perturbation 1 est rejetée avec une p-value inférieure à 2.2×10^{-16} (voir Annexe D pour le détail du test de Kolmogorov-Smirnov). Cette perturbation n'est donc pas adaptée à notre problématique. De plus, perturber directement le log ratio sans tenir compte des valeurs des intensités occulte les bruits sur le log ratio importants observés pour les faibles intensités. Les perturbations testées par la suite prennent donc en compte les valeurs d'intensité.

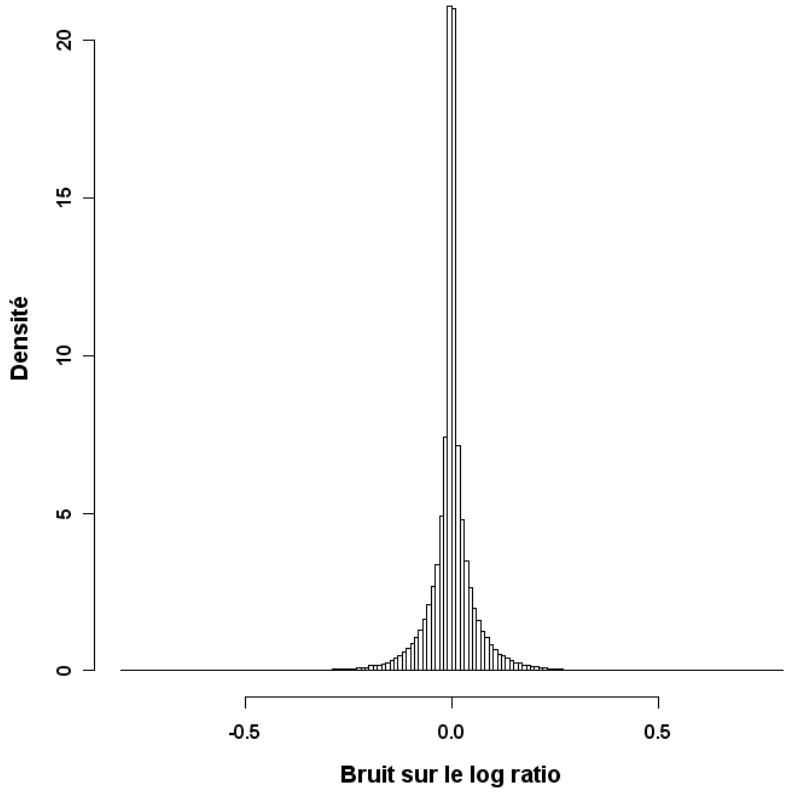

FIG. 3.9 : Histogramme du bruit sur le log ratio pour la perturbation 1
Les barres sont de largeur 0.01.

3.2.3.2 Perturbation 2 : perturber les intensités sans lien entre intensité Cy3 et intensité Cy5

Dans un premier temps, les intensités ont été perturbées de manière indépendante. Un nouveau log ratio était en suite calculé. Cette étape a été réalisée à partir de l'observation du bruit sur l'intensité en fonction de l'intensité (Figure 3.7.a). Chaque intensité (Cy3 et Cy5) a

été perturbée par l'ajout d'une variable aléatoire de moyenne nulle et d'écart type $0.124\times I$, où I est la valeur de l'intensité considérée. Cet écart type a été estimé à partir de la relation entre écart type de l'intensité sur les réplicats techniques et moyenne de l'intensité grâce à une droite de régression (Figure 3.10). Trois lois ont été testées pour cette variable aléatoire : loi uniforme sur l'intervalle $[-\sqrt{3}\times 0.124\times I, \sqrt{3}\times 0.24\times I]$ (type de perturbation proposé par Sayyed-Ahmad), loi normale d'écart type $0.124\times I$ et loi normale au carré signée d'écart type $\sqrt{\frac{0.124\times I}{\sqrt{3}}}$.

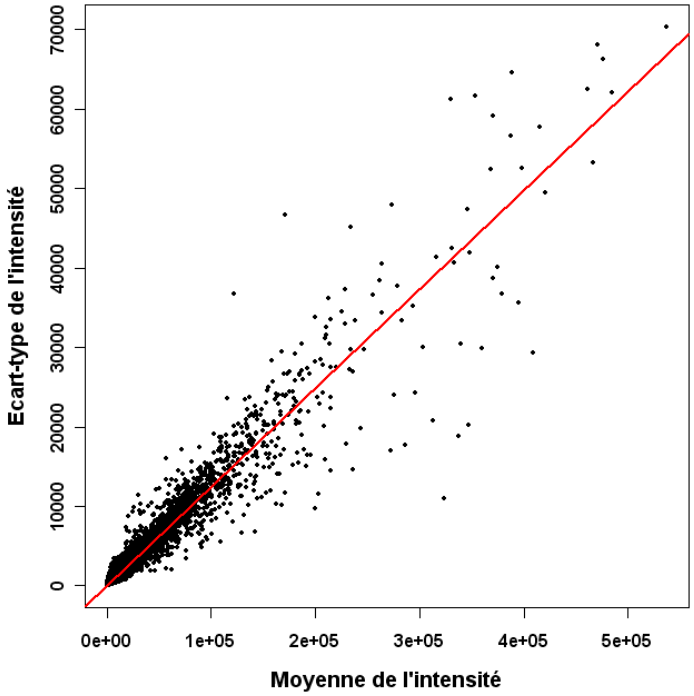

FIG. 3.10 : Écart type de l'intensité sur les différents réplicats techniques en fonction de la moyenne de l'intensité sur ces mêmes réplicats
La droite représentée en rouge est la droite de régression.

Les résultats présentés ici sont ceux de la loi ayant reproduit le mieux la relation entre intensité et bruit sur l'intensité, c'est-à-dire la loi normale (Figure 3.11.a). Cependant, le bruit final obtenu sur le log ratio a une distribution trop large qui n'est pas du tout cohérente avec la réalité (Figure 3.11.b). Par la suite, le lien entre les bruits sur les deux intensités a donc été pris en compte.

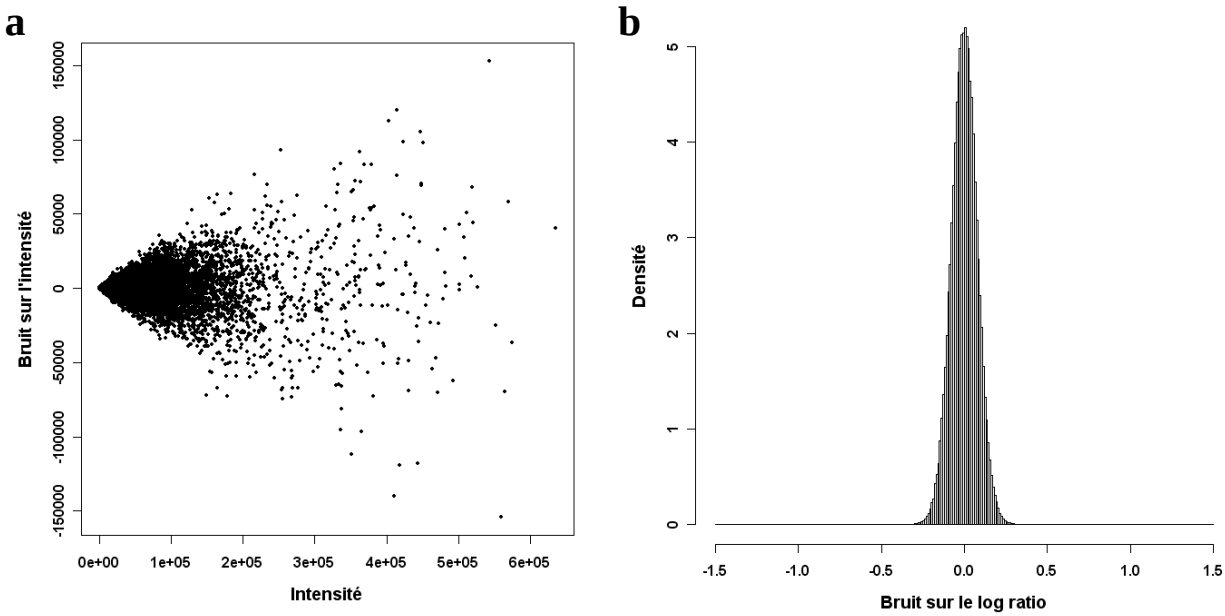

FIG. 3.11 : Résultats obtenus avec la perturbation 2
(a) : Bruit sur l'intensité en fonction de l'intensité. (b) : Histogramme du bruit sur le log ratio. Les barres sont de largeur 0.01.

3.2.3.3 Perturbation 3 : perturber les intensités avec lien entre intensité Cy3 et intensité Cy5

Pour la perturbation 3, les bruits ont été choisis afin d'essayer de reproduire le lien entre bruit sur l'intensité Cy3 et bruit sur l'intensité Cy5 tout en respectant ce qui a déjà été observé dans la perturbation précédente. On observe sur la Figure 3.7.b que les bruits sur les deux intensités semblent liés par une relation linéaire à une variable aléatoire près, que l'on note V, dont l'écart type est plus important pour les bruits élevés. Il a donc été décidé de choisir un bruit moyen BM à partir de la moyenne des intensités Cy3 et Cy5 comme expliqué pour la perturbation 2 (loi normale d'écart type $0.124 \times \mathrm{I}$), puis d'obtenir les bruits sur les intensités Cy3 (verte) et Cy5 (rouge) de la manière suivante : $\mathrm{BIV} = \mathrm{BM} - \frac{\mathrm{V}}{2}$ et $\mathrm{BIR} = \mathrm{BM} + \frac{\mathrm{V}}{2}$. V reste à déterminer.

Afin d'étudier plus précisément les propriétés de la différence entre les bruits sur les deux intensités, cette différence a été représentée en fonction de l'intensité moyenne (Figure 3.12.a). On observe une allure assez similaire à la Figure 3.7.a mais en plus condensé, avec un écart type croissant avec l'intensité. La variable aléatoire V a dont été définie comme suivant une loi normale au carré signée, centrée en zéro et d'écart type dépendant de la moyenne des intensités. D'après la représentation de cet écart type en fonction de l'intensité moyenne

(Figure 3.12.b), il est difficile de trouver une relation exacte. Néanmoins, le problème a été simplifié en approchant la relation entre écart type de V et moyenne des intensités par une droite d'équation : $y = 0.071 \times x$. V suit donc une loi normale au carré signée centrée en zéro et d'écart type $\sqrt{\frac{0.071 \times I}{\sqrt{3}}}$.

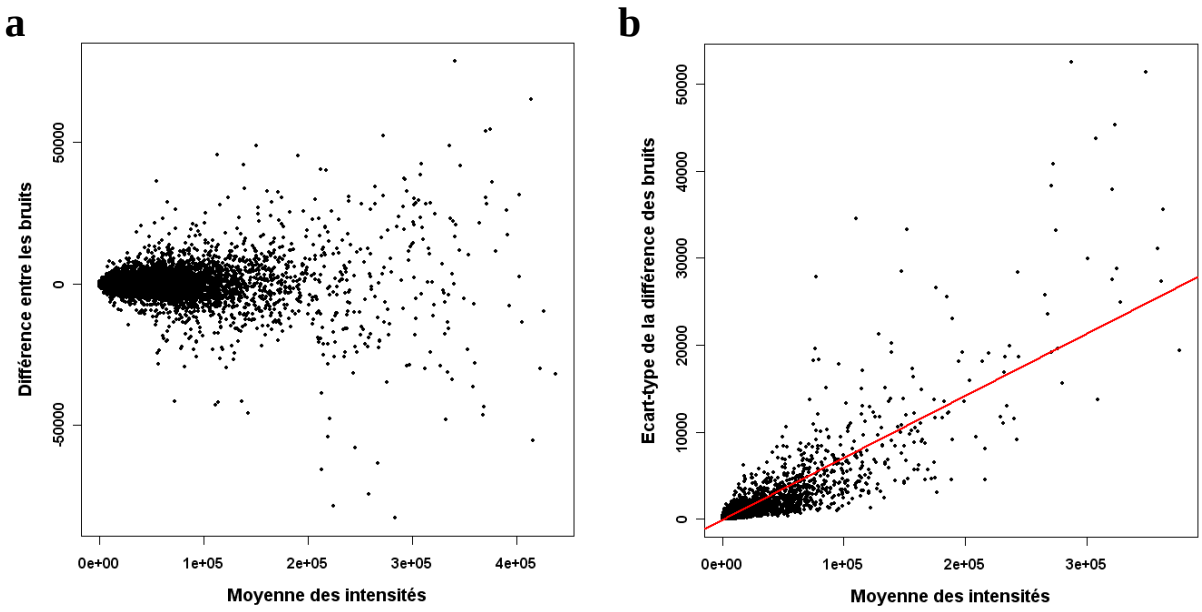

FIG. 3.12 : Étude de la différence entre bruit sur l'intensité Cy5 et bruit sur l'intensité Cy3
(a) : Différence entre les bruits sur les intensités Cy5 et Cy3 en fonction de la moyenne des intensités Cy5 et Cy3. (b) : Écart type de la différence entre les bruits sur les intensités Cy5 et Cy3 en fonction de la moyenne des intensités Cy5 et Cy3.

Après application de cette perturbation, le graphe de la différence entre les bruits des intensités Cy5 et Cy3 en fonction de la moyenne des intensités n'est pas tout à fait équivalent à celui obtenu sur données réelles (Figure 3.13.a). Cette observation n'est pas étonnante au vu des approximations qui ont été faites. La distribution du bruit sur le log ratio obtenue au final est par contre assez proche de l'originale, mais toujours pas suffisamment (Figure 3.13.b). En effet, un test de Kolmogorov-Smirnov rejette l'hypothèse de même distribution initiale pour le bruit réel et le bruit de la perturbation 3 avec une p-value inférieure à 2.2×10^{-16}.

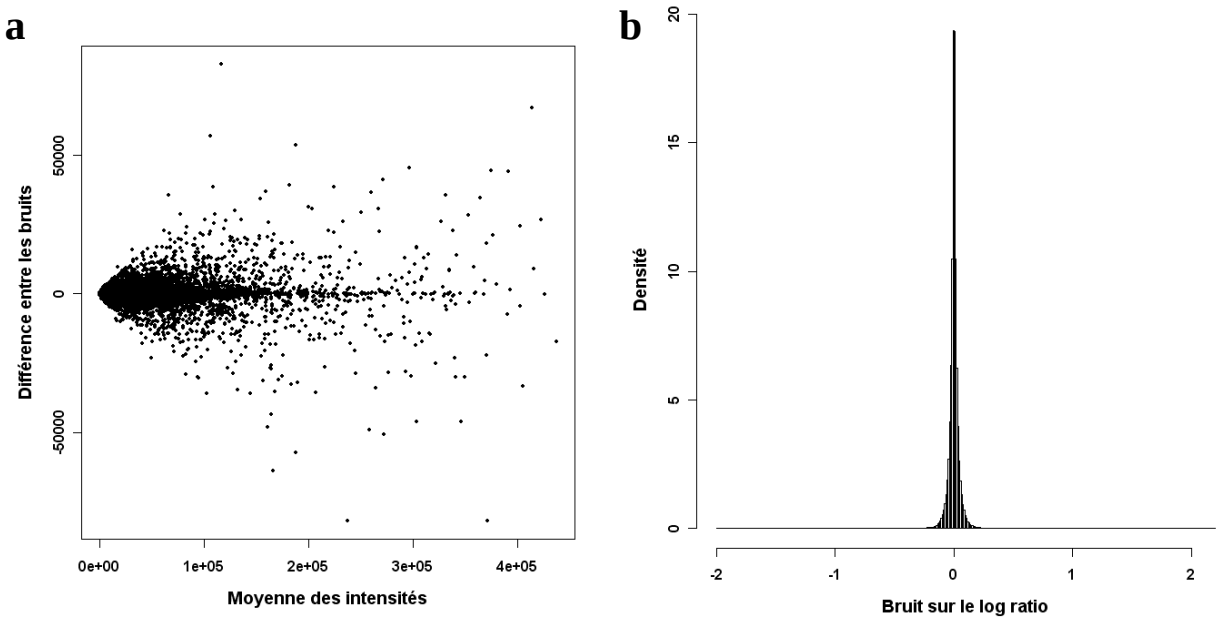

FIG. 3.13 : Résultats obtenus avec la perturbation 3
(a) : Différence entre les bruits sur les intensités Cy5 et Cy3 en fonction de la moyenne des intensités Cy5 et Cy3. (b) : Histogramme du bruit sur le log ratio. Les barres sont de largeur 0.01.

3.2.3.4 Perturbation 4 : perturber le log ratio en fonction des intensités

La difficulté de reproduire le bruit technique à partir de lois de probabilités a conduit à s'interroger sur la possibilité d'une perturbation empirique des données prenant en compte les observations déjà effectuées. On souhaite à la base perturber le log ratio. Cependant, les résultats précédents ont plutôt suggéré une influence de la valeur des intensités sur le bruit. Il a donc été décidé de s'inspirer du graphe de la Figure 3.8.a, sur lequel on observe une variation de l'étalement du bruit sur le log ratio en fonction de la moyenne des intensités Cy5 et Cy3.

Le principe utilisé est donc le suivant. Pour une séquence dont on cherche à perturber le log ratio, un bruit est tiré au hasard parmi les bruits réels trouvés pendant l'étude de variabilité technique. Cependant, ce bruit n'est pas tiré parmi tous les bruits réels, mais parmi ceux qui correspondent sur la Figure 3.8.a à un intervalle d'intensité contenant la moyenne des intensités Cy5 et Cy3 de la séquence considérée. Ce bruit est ensuite ajouté à la valeur initiale du log ratio.

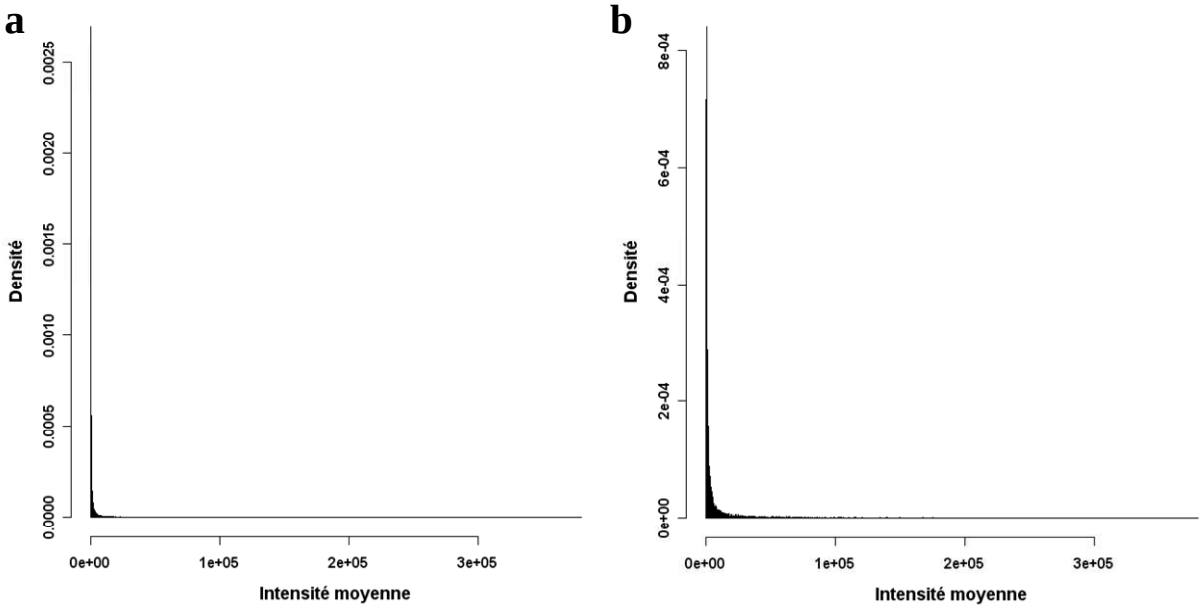

FIG. 3.14 : Histogramme des intensités moyennes
(a) : Toutes les valeurs d'intensité. (b) : La moitié supérieure des valeurs d'intensité supérieures. Les barres sont de largeur 100.

En pratique, tous les bruits de tous les réplicats techniques obtenus lors de l'étude de variabilité technique ont été regroupés et ordonnés en fonction de la moyenne des intensités leur correspondant (moyenne sur les réplicats techniques et moyenne sur les intensités Cy3 et Cy5). Comme le nombre de valeurs disponibles décroît exponentiellement avec l'intensité (Figure 3.14a et Figure 3.14b), les intervalles d'intensité ont été définis de la manière suivante. Ils sont de longueurs croissantes valant la suite des puissances de 10 (L dans {1,10,100,...,100000}) avec une exception pour le premier intervalle qui contient également les intensités très faibles ne disposant pas de données de bruit dans l'étude de variabilité technique. La longueur de l'intervalle est augmentée quand il n'y a pas au moins 18 points dans un intervalle. Ce nombre correspond à un minimum de trois séquences avec six réplicats techniques par intervalle.

Une telle perturbation permet effectivement de retrouver l'allure du graphe du bruit sur le log ratio en fonction de la moyenne des intensités Cy3 et Cy5 (Figure 3.15.a), mais également d'obtenir une excellente distribution du bruit sur le log ratio (Figure 3.15.b). Un test de Kolmogorov-Smirnov fournit une p-value non significative de 0.43. L'hypothèse nulle de provenance d'une même distribution pour les données de bruit réelles et les données de bruit de la perturbation 4 n'est donc pas rejetée.

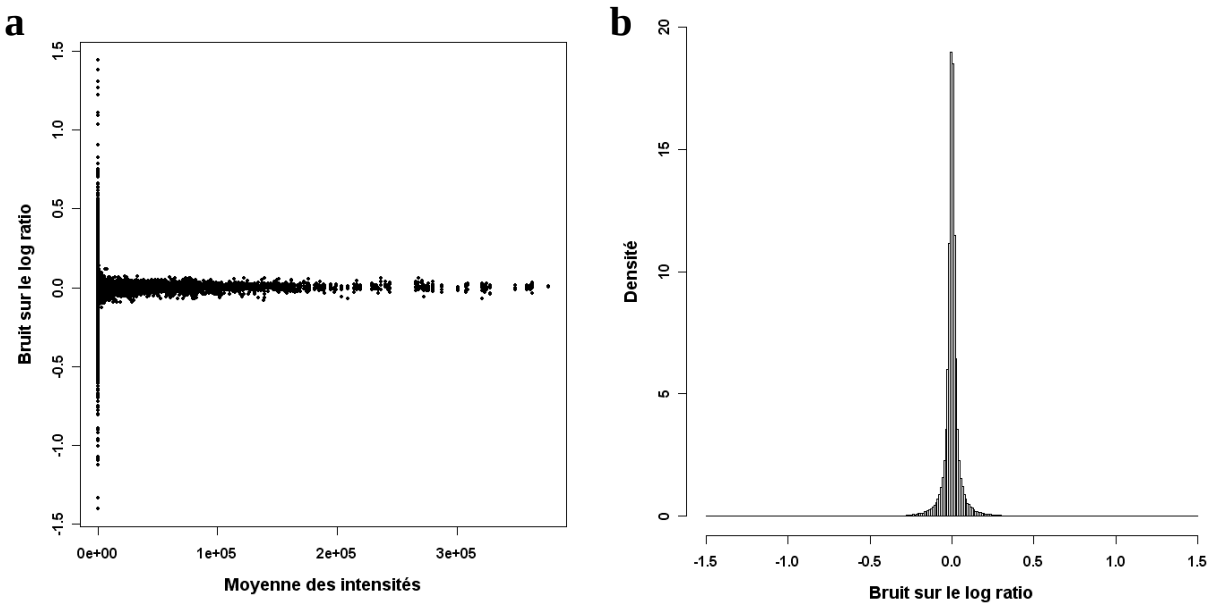

FIG. 3.15 : Résultats obtenus avec la perturbation 4
(a) : Bruit sur le log ratio en fonction de la moyenne des intensités Cy3 et Cy5. (b) : Histogramme du bruit sur le log ratio. Les barres sont de largeur 0.01.

3.2.4 Conclusion

Dans l'espoir d'obtenir une forme de perturbation générique qui puisse être réutilisée pour d'autres jeux de données, des perturbations utilisant des lois de probabilités ont tout d'abord été testées. Parmi celles-ci, les perturbations 1 et 3 ont obtenu les distributions de bruit sur le log ratio se rapprochant visuellement le plus de la réalité. Cependant, des écarts demeurent encore et une perturbation empirique, la perturbation 4, a donc été testée. C'est la perturbation donnant les résultats les plus proches des données réelles. Elle a donc été choisie pour être utilisée dans MetRob. L'étape suivante consiste à déterminer les valeurs à attribuer aux paramètres des différentes méthodes de sélection de variables utilisées (T-test, NSC et SVM-RFE).

3.3 Paramètres des méthodes de sélection de variables

3.3.1 T-test

Le T-test avec correction de Welch est un test statistique qui permet de vérifier la significativité de la différence entre les moyennes de deux populations dans le cas où les variances sont différentes. C'est habituellement la p-value du test qui détermine les séquences significatives. La p-value du T-test représente la proportion de faux positifs rapportée aux séquences réellement non significatives. Dans le cadre d'une application aux puces à ADN, le nombre de variables est très important et le nombre de variables non significatives également. Une p-value de 0.05 peut alors représenter beaucoup de faux positifs. Elle doit donc être corrigée par une procédure adaptée aux tests multiples.

Classement	q-value	Classement	q-value	Classement	q-value
1	0.212655636	11	0.920395665	21	0.999525061
2	0.212655636	12	0.987112805	22	0.999525061
3	0.212655636	13	0.987112805	23	0.999525061
4	0.212655636	14	0.987112805	24	0.999525061
5	0.212655636	15	0.987112805	25	0.999525061
6	0.378007517	16	0.987112805	26	0.999525061
7	0.378007517	17	0.999525061	27	0.999525061
8	0.527107242	18	0.999525061	28	0.999525061
9	0.920395665	19	0.999525061	29	0.999525061
10	0.920395665	20	0.999525061	30	0.999525061

TAB. 3.1 : 30 premières valeurs de q-value obtenues sur le jeu de données RS_TAI_dose1 et classées par ordre croissant

Une solution couramment utilisée est le calcul d'une q-value. Cette q-value représente le taux de fausse découverte, c'est-à-dire la proportion de faux positifs rapportée aux séquences réellement significatives. Des q-values ont donc été calculées via le package qvalue de R [69]. Il s'est néanmoins avéré que la correction appliquée était trop importante pour permettre de classer correctement les séquences sur tous les jeux de données. En effet, pour le jeu de données RS_TAI_dose1 correspondant à la dose 28 µmol/kg dans le tissu adipeux inguinal, les valeurs de q-value obtenues ne sont significatives pour aucune séquence et sont égales à 0.999525061 de la $17^{ème}$ à la $10247^{ème}$ séquence (Tableau 3.1). Il est donc impossible d'obtenir un classement pertinent. Ces résultats peuvent être dus à de faibles différences entre les classes associées à un petit nombre d'observations. Comme seul un classement des séquences était recherché, le problème des tests multiples a été négligé et la p-value classique du T-test a été conservée. Par ailleurs, aucun seuil de p-value n'a été nécessaire puisque toutes les séquences étaient conservées à ce stade.

3.3.2 Nearest Shrunken Centroids

La méthode de Nearest Shrunken Centroids (NSC) ne nécessite qu'un seul paramètre, le niveau de contraction Δ, qui détermine les séquences à considérer comme significatives. Comme NSC n'a été utilisée que pour classer les séquences, Δ a été fixé à zéro.

3.3.3 Support Vector Machines – Recursive Feature Elimination

Le package Spider de Matlab a été utilisé pour la Support Vector Machines – Recursive Feature Elimination (SVM-RFE). Les paramètres modifiables dans ce package sont :

- kernel : type de noyau utilisé pour le SVM
- C : paramètre de soft-margin qui pondère le poids accordé aux exemples mal classés
- ridge : paramètre permettant d'éviter les matrices non inversibles (sa valeur par défaut est de 10^{-13} et a été conservée)
- feat : nombre de variables à conserver au final (ce paramètre a uniquement servi pour raccourcir le temps de calcul des tests, mais il a été fixé à 1 par la suite pour classer toutes les séquences)
- speed : nombre de variables à partir duquel on ne retire qu'une seule variable à la fois (avant le nombre de variables est divisé par deux à chaque itération)

Trois jeux de données test ont été utilisés pour choisir ces paramètres : RS_Soleus_dose1, RS_Foie_dose2 et RS_TAI_dose3.

3.3.3.1 Choix du noyau

Trois types de noyaux ont été testés pour le SVM : linéaire, polynomial (de degrés 2, 3, 4 et 5) et gaussien (d'écart type σ valant 0.1, 0.5, 1, 2 et 10). Les autres paramètres étaient fixés aux valeurs suivantes : $C = +\infty$, speed = 1000, $feat \in \{100, 500\}$. La qualité des résultats a été évaluée par l'erreur obtenue en Leave-One-Out-Cross-Validation (LOOCV) et est présentée dans le Tableau 3.2. C'est avec le noyau linéaire que l'on obtient globalement les meilleurs résultats, la moitié des erreurs par LOOCV étant nulles. En revanche, quel que soit le degré utilisé, le noyau polynomial n'est pas très performant. L'erreur obtenue avec le noyau gaussien est faible uniquement pour un écart type à 10 et reste plus élevée qu'avec le noyau linéaire. Le noyau linéaire a donc été choisi pour ses résultats et sa simplicité.

Noyau	feat = 100			feat = 500		
	RS_Soleus_ dose1	RS_Foie_ dose2	RS_TAI_ dose3	RS_Soleus_ dose1	RS_Foie_ dose2	RS_TAI_ dose3
NL	0.083	0	0.167	0	0	0.083
NP degré 2	0.417	0.917	0.5	0.417	0.917	0.5
NP degré 3	0.417	0.917	0.917	0.25	1	0.833
NP degré 4	0.5	0.917	0.583	0.5	0.917	0.667
NP degré 5	0.583	0.5	0.5	0.167	0.583	0.25
NG s 0.1	1	1	1	1	1	1
NG s 0.5	1	1	1	1	1	1
NG s 1	0.5	1	1	1	1	1
NG s 2	0.25	1	1	1	1	1
NG s 10	0	0.167	0.25	0.083	0.083	0.417

TAB. 3.2 : Erreurs par LOOCV obtenues avec différents types de noyaux
NL : noyau linéaire, NP : noyau polynomial, NG : noyau gaussien

3.3.3.2 Choix de C, paramètre de soft-margin

Différentes valeurs du paramètre de soft-margin C ont été testées afin d'étudier son impact sur chaque type de noyau : $C \in \{+\infty, 100, 10, 1, 0.1\}$. C correspond au poids attribué aux exemples mal classés. Ainsi, la valeur de C par défaut, égale à +∞, interdit les mauvaises classifications sur les données d'apprentissage. Un noyau linéaire, un noyau polynomial de degré 2 et un noyau gaussien d'écart type 1 ont été utilisés. Le paramètre feat a pris les valeurs 100 et 500, speed était fixé à 1000 et les performances ont été évaluées avec l'erreur par LOOCV. Les résultats sont uniquement présentés pour le noyau linéaire car les conclusions sont identiques pour les autres noyaux (Tableau 3.3).

C	feat = 100			feat = 500		
	RS_Soleus_ dose1	RS_Foie_ dose2	RS_TAI_ dose3	RS_Soleus_ dose1	RS_Foie_ dose2	RS_TAI_ dose3
Infini	0.083	0	0.167	0	0	0.083
100	0.083	0	0.083	0	0	0.083
10	0.083	0	0.167	0	0	0.083
1	0.083	0	0.167	0	0	0.083
0,1	0.083	0	0.083	0	0	0.083

TAB. 3.3 : Erreurs par LOOCV obtenues avec différentes valeurs de C pour chaque type de noyau

Modifier le paramètre C ne change presque pas les performances de classification du SVM. Il semble donc pertinent de choisir la valeur de C la plus restrictive sur la qualité de la classification, c'est-à-dire $C = +\infty$. Les listes de séquences obtenues avec les différentes valeurs de C ont donc été comparées à celle obtenue avec $C = +\infty$ afin de vérifier que les listes n'étaient pas trop différentes (Tableau 3.4). Pour $feat = 500$, on observe assez peu de différences entre $C = +\infty$ et les autres valeurs. Ces différences sont plus importantes pour $feat = 100$. Comme feat sera égal au nombre total de séquences pour la méthodologie finale, la valeur de $C = +\infty$ a été conservée. On peut noter que, lors du calcul de l'erreur par LOOCV, Spider construit une liste de séquence par sous-ensemble d'apprentissage. Ces listes peuvent donc différer de la liste de séquences finale fournit. Ceci explique que deux valeurs de C peuvent fournir des erreurs différentes mais des listes de séquences finales identiques.

	feat = 100			**feat = 500**		
C	**RS_Soleus_ dose1**	**RS_Foie_ dose2**	**RS_TAI_ dose3**	**RS_Soleus_ dose1**	**RS_Foie_ dose2**	**RS_TAI_ dose3**
100	85	89	91	98.6	92.2	96.8
10	87	94	93	98.6	100	97.4
1	100	94	100	100	100	100
0,1	100	94	100	100	100	100

TAB. 3.4 : Pourcentage de séquences communes entre les listes obtenues avec les différentes valeurs de C et celle obtenue avec $= +\infty$

3.3.3.3 Choix de speed, paramètre de vitesse

En dernier lieu, le paramètre speed de la RFE a été considéré. Avec un noyau linéaire, $C = +\infty$ et feat valant 100 et 500, quatre valeurs de speed ont été testées. Pour chaque jeu de données test, ces valeurs ont été choisies de manière à diviser le nombre de séquences par deux, 0 fois, 1 fois, 2 fois ou 3 fois. Les erreurs par LOOCV obtenues sont présentées dans le Tableau 3.5 et montrent que les modifications de speed ont peu d'impact sur les performances de classification.

	feat = 100			**feat = 500**		
speed	**RS_Soleus_ dose1**	**RS_Foie_ dose2**	**RS_TAI_ dose3**	**RS_Soleus_ dose1**	**RS_Foie_ dose2**	**RS_TAI_ dose3**
division 0 fois	0	0	0.167	0	0	0.167
division 1 fois	0	0	0.167	0	0	0.167
division 2 fois	0	0	0.167	0	0	0.167
division 3 fois	0.083	0	0.167	0	0	0.167

TAB. 3.5 : Erreurs par LOOCV obtenues avec différentes valeurs de speed

La liste de séquences la plus pertinente est censée être obtenue avec une valeur de speed telle que le nombre de séquences ne soit jamais divisé par deux. En effet, les séquences sont alors toutes retirées une par une et le classement final est plus précis. Cependant, cette précision est coûteuse en temps de calcul. Il est donc intéressant de pouvoir choisir une vitesse plus élevée sans modifier les résultats. Les listes de séquences obtenues avec les différentes valeurs de speed ont donc été comparées à celle obtenue avec la valeur de speed correspondant à zéro division (Tableau 3.6). Seule la valeur de speed correspondant à diviser 3 fois les séquences par deux semble modifier les résultats. Afin de rester cohérent avec les observations faites sur les anciennes lames 22k (voir Annexe C) et de s'assurer un bon classement de la première moitié des séquences, la valeur de speed divisant le nombre de variables par deux une seule fois a été choisie.

	feat = 100			**feat = 500**		
speed	**RS_Soleus_ dose1**	**RS_Foie_ dose2**	**RS_TAI_ dose3**	**RS_Soleus_ dose1**	**RS_Foie_ dose2**	**RS_TAI_ dose3**
division 1 fois	100	100	100	100	100	100
division 2 fois	100	100	100	100	100	100
division 3 fois	87	92	100	90	98.6	100

TAB. 3.6 : Pourcentage de séquences communes entre les listes obtenues avec les différentes valeurs de speed et celle obtenue avec la valeur de speed correspondant à zéro division par deux du nombre de séquences

Ainsi, la méthode de SVM-RFE a été utilisée dans notre étude avec les paramètres suivants :

- kernel = noyau linéaire
- feat = 1
- speed = nombre total de variables − 1
- C = +∞
- ridge = 1^{e} − 13

3.4 Paramètres de MetRob

Différentes décisions ont été prises concernant la méthodologie MetRob : modalité de choix d'un nombre de séquences, nombre de perturbations, seuil de reproductibilité. Elles sont justifiées dans les parties suivantes. Les mêmes jeux de données test que précédemment ont été conservés.

3.4.1 Modalité de choix d'un nombre de séquences

Dans MetRob, les méthodes T-test, NSC et SVM-RFE sont seulement utilisées pour classer les séquences : par p-values pour le T-test, par scores pour NSC et par ordre inverse d'élimination pour SVM-RFE. Il faut ensuite déterminer le nombre de séquences les plus discriminantes. Pour cela, on cherche à maximiser la robustesse de la liste de séquences choisie tout en minimisant la longueur de cette liste. Pour rappel, la robustesse est définie comme le pourcentage moyen de séquences communes entre une liste de séquences obtenue sur les données réelles et des listes de séquences de même longueur obtenues sur des données perturbées.

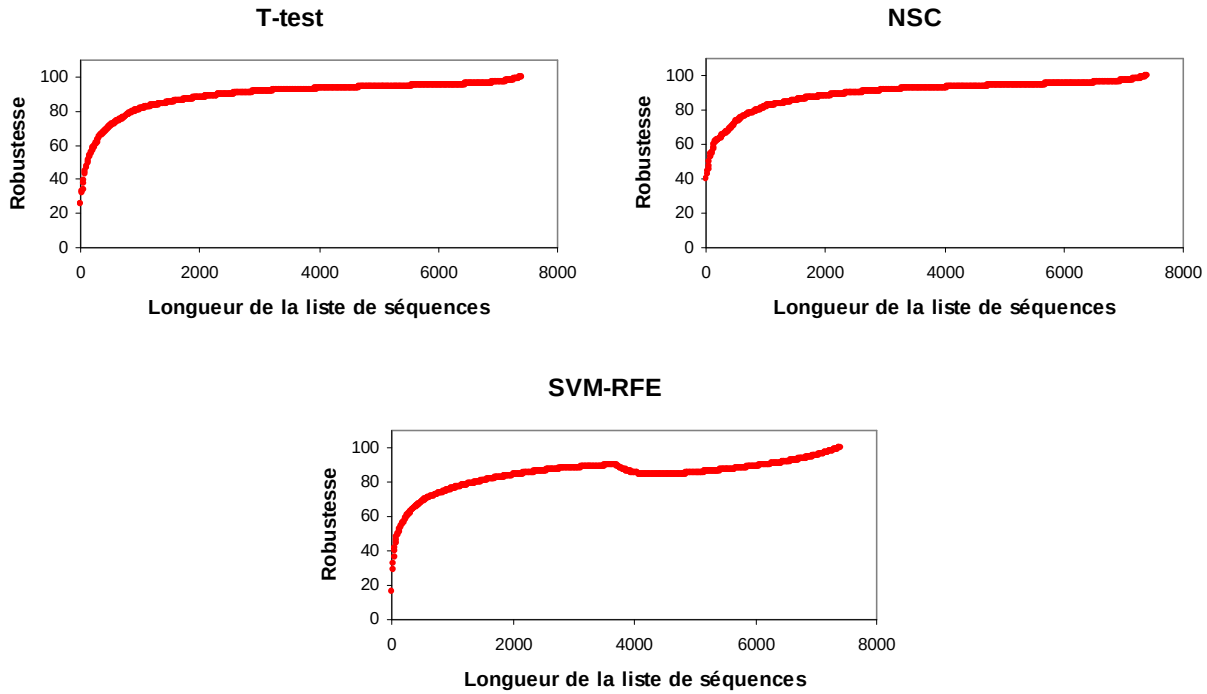

FIG. 3.16 : Graphes de la robustesse en fonction du nombre de séquences considérées pour les trois méthodes T-test, NSC et SVM-RFE
Cas du jeu de données RS_Soleus_dose1 pour 300 perturbations.

La robustesse a été calculée dans notre étude pour des longueurs L de liste de séquences allant de 10 en 10, $L \in \{10,20,30,\ldots,N\}$, où N est le plus grand multiple de 10 inférieur au nombre total de séquences. Elle a été représentée en fonction de la longueur de la liste de séquences pour chaque méthode et pour le jeu de données RS_Soleus_dose1 avec 300 perturbations (Figure 3.16). On observe que la robustesse croît avec le nombre de séquences jusqu'à atteindre 100% quand toutes les séquences sont conservées. D'autre part, les pentes de ces courbes décroissent globalement sur la première moitié des graphes, ce qui signifie que les séquences rajoutées améliorent de moins en moins la robustesse. Pour la méthode SVM-RFE,

on peut remarquer un point d'inflexion correspondant à la moitié des séquences. Il s'agit en fait d'un artefact lié au choix de la vitesse de la RFE : la dernière moitié des séquences n'a pas vraiment été classée.

Dans l'optique de la recherche d'un nombre optimal de séquences, on définit : $\mathrm{Diff(L)} = \frac{\mathrm{Rob(L)}}{100} - \frac{\mathrm{L}}{\mathrm{NbSeqTot}}$, où Rob(L) est la robustesse associée à une longueur L de liste de séquences. NbSeqTot est le nombre total de séquences statistiquement significativement régulées dans l'étude. Maximiser la robustesse en conservant un nombre de séquences minimal correspond alors à trouver la longueur de liste de séquences qui maximise Diff. Avec la méthode NSC, la fonction Diff a l'allure présentée sur la Figure 3.17 pour les jeux de données RS_Soleus_dose1 et RS_TAI_dose1. La majorité des jeux de données engendre une courbe comme celle de RS_Soleus_dose1, c'est-à-dire une courbe présentant visuellement un maximum correspondant à une annulation de la pente. Cependant, on observe que pour le jeu de données RS_TAI_dose1, le vrai maximum est en fait le premier point de la courbe calculé pour 10 séquences (point entouré d'un cercle noir).

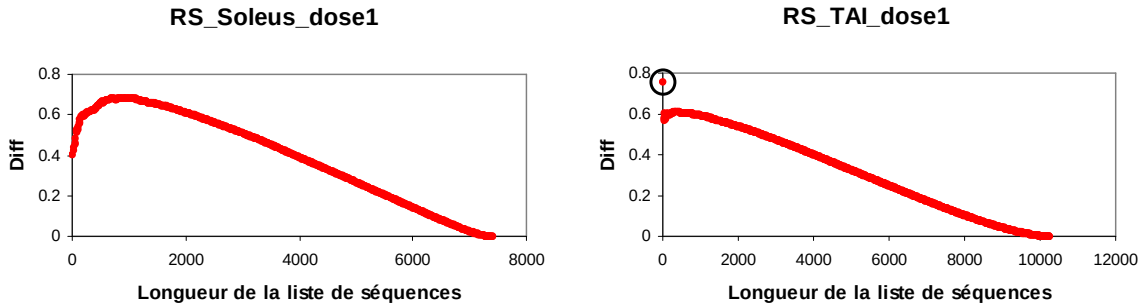

FIG. 3.17 : Graphe de Diff en fonction du nombre de séquences pour les jeux de données RS_soleus_dose1 et RS_TAI_dose1
Cas de la méthode NSC

En conséquence, le maximum de Diff a été principalement évalué en recherchant le point d'annulation de la pente. La première étape consiste à calculer les pentes S_{iq} de la courbe Diff(L) pour chaque point i à des ordres q croissants : $S_{iq} = \frac{\mathrm{Diff}(i+10\times q) - \mathrm{Diff}(i-10\times q)}{2q\times 10}$.

Le plus petit ordre pour lequel la pente est toujours positive puis toujours négative est sélectionné, puis la plus grande longueur de liste de séquences donnant une pente positive à cet ordre est choisie. Les ordres q croissants permettent de lisser la pente et d'être plus reproductible sur différents lancements de perturbations aléatoires. La seconde étape gère les situations particulières où le maximum est au tout début de la courbe (ex : RS_TAI_dose1) ou pour lesquelles la courbe est très irrégulière. Le maximum empirique de Diff est sélectionné. Si ce maximum dépasse de plus de 1% la valeur choisie pendant la première étape, la longueur de liste de séquences associée à ce maximum empirique est choisie.

3.4.2 Choix d'un nombre de perturbations

Dans la méthodologie MetRob, les données sont perturbées plusieurs fois pour compenser le caractère aléatoire des perturbations. Les résultats sont ensuite moyennés afin de calculer la robustesse pour plusieurs nombres de séquences, puis d'en déduire une longueur optimale de liste de séquences. Il faut donc déterminer le nombre de perturbations à utiliser et différentes valeurs ont été testées. Les longueurs de listes de séquences obtenues avec N perturbations ont été comparées à celle, considérée arbitrairement comme idéale et notée L, obtenue avec 1500 perturbations (nombre total maximal de perturbations lancées). Les calculs ont été lancés plusieurs fois afin de pallier le caractère aléatoire des perturbations.

On note L_{Nk}, la longueur de liste de séquences obtenue au $k^{ième}$ lancement de N perturbations. Pour m lancements de N perturbations, on définit $C(N,m) = \underset{k\in\{1,\ldots,m\}}{\text{Moyenne}}\left(\left|L_{Nk} - L\right|\right)$. Ce critère a été utilisé pour évaluer la qualité des résultats en recherchant le plus petit C(N,m). Les valeurs $(N,m) \in \{(100,15),(200,7),(300,5),(400,3),(500,3)\}$ ont été testées (Figure 3.18). On peut noter que la valeur $N = 500$ est déjà difficile à utiliser en pratique : 50 heures de calcul pour la SVM-RFE avec seulement 8000 séquences. Elle est présentée pour compléter les résultats mais a été exclue des bonnes solutions.

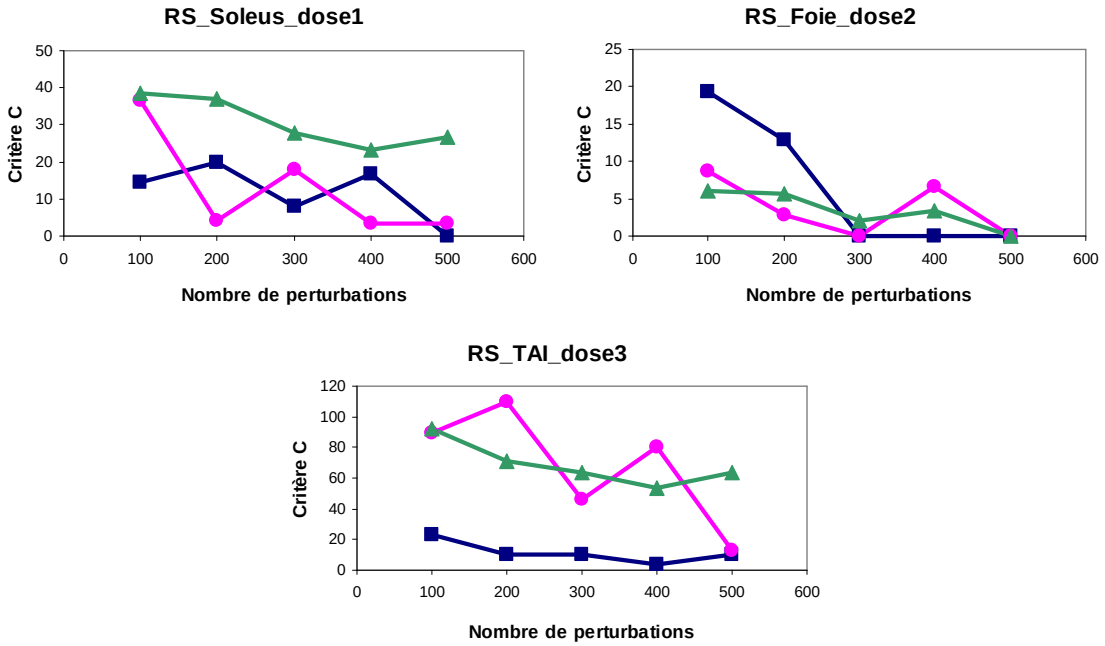

FIG. 3.18 : Graphes du critère C en fonction du nombre de perturbations N
Les carrés bleus (■) correspondent au T-test, les cercles roses (●) à NSC et les triangles verts (▲) à la SVM-RFE. A noter que les échelles sont différentes entre les graphes.

Pour le jeu de données RS_Soleus_dose1, le nombre de perturbations donnant la meilleure valeur de C dépend de la méthode de sélection de variables utilisée. Le choix se porte sur N = 400 pour le T-test et la SVM-RFE et sur N = 300 pour NSC. Sur les données de RS_Foie_dose2, les trois méthodes s'accordent pour une valeur de N = 300. Enfin, le dernier jeu de données, RS_TAI_dose3, fournit des valeurs optimales pour N = 400 avec le T-test et la SVM-RFE et pour N = 300 avec NSC. Il est donc difficile de conclure car les résultats dépendent du jeu de données, même pour une seule méthode. La valeur de N = 300 a finalement été utilisée pour des raisons de temps de calcul.

3.4.3 Choix d'un seuil de reproductibilité

La liste de séquences obtenue après avoir choisi une longueur de liste adéquate ne peut pas directement être considérée comme la liste finale de MetRob. En effet, le pourcentage de séquences trouvées dans tous les jeux de données (jeu de données non perturbées et 300 jeux de données perturbées) est très faible (entre 5 et 16% : Tableau 3.7). Cela signifie que peu de séquences peuvent être sélectionnées quelle que soit la perturbation et que la liste n'est pas vraiment robuste. Ceci souligne également l'importance de prendre en compte l'impact de la variabilité technique sur les résultats.

		Nombre de séquences optimal	**Intersection sur toutes les perturbations**	**Pourcentage**
RS_Soleus_dose1	**T-test**	970	128	**13.2**
	NSC	970	152	**15.67**
	SVM-RFE	830	44	**5.3**
RS_Foie_dose2	**T-test**	260	36	**13.85**
	NSC	250	30	**12**
	SVM-RFE	270	31	**11.48**
RS_TAI_dose3	**T-test**	1670	109	**6.53**
	NSC	1510	104	**6.89**
	SVM-RFE	1150	92	**8**

TAB. 3.7 : Non-reproductibilité au travers des perturbations
L'intersection sur toutes les perturbations est le nombre de séquences sélectionnées avec les données réelles et avec tous les jeux de données perturbées

Une étape supplémentaire est donc nécessaire. Il s'agit de l'étape consistant à choisir parmi cette liste de séquences, celles qui sont présentes dans au moins $p = 80\%$ des résultats sur données perturbées. Cette valeur de p, pourcentage des jeux de données perturbées pour lesquels une séquence doit avoir été trouvée afin d'être sélectionnée, a été choisie de manière à optimiser la reproductibilité de la liste de séquences. Par reproductible, on entend que les listes de séquences obtenues sur différents lancements de 300 perturbations doivent être proches (Figure 3.19).

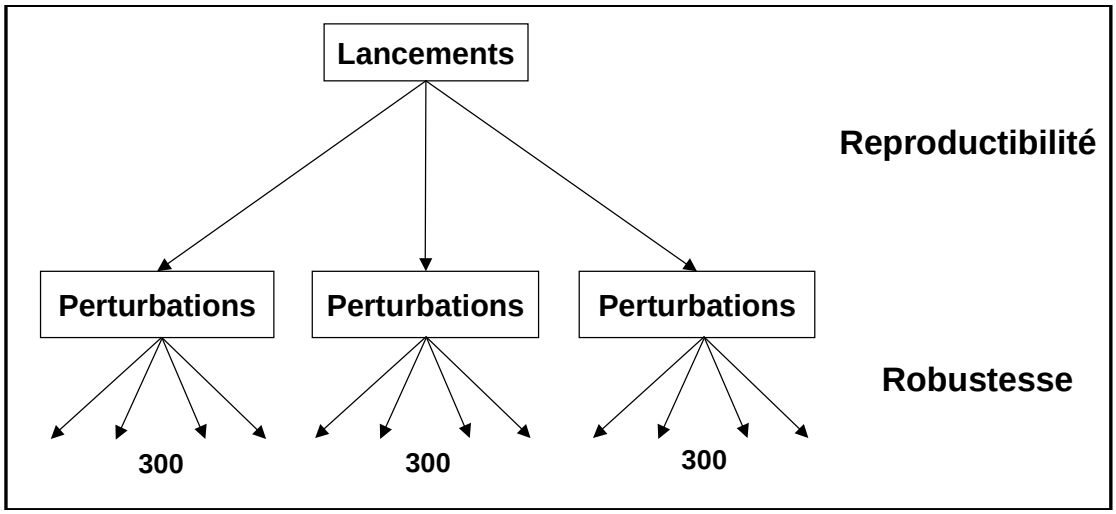

FIG. 3.19 : Reproductibilité et robustesse
La reproductibilité représente la ressemblance entre des listes de séquences obtenues pour des lancements différents. La robustesse représente la ressemblance entre des listes de séquences obtenues pour un jeu de données non perturbées et pour un jeu de données perturbées.

Dans un premier temps, seules les séquences présentes dans $p = 100\%$ des résultats sur données perturbées avaient été sélectionnées, mais les résultats n'étaient pas reproductibles. Avec $p = 100\%$, si une séquence est absente pour une seule perturbation, cela est suffisant pour la supprimer de la liste de séquences finale. Comme les perturbations sont aléatoires, il est alors très facile d'avoir des résultats différents sur plusieurs lancements. La Figure 3.20 présente les résultats en terme de reproductibilité avec différentes valeurs de p ($p \in \{50, 75, 80, 90, 95, 99, 100\}$). Les calculs ont été lancés trois fois avec 300 perturbations et le pourcentage de séquences trouvées dans les trois lancements, c'est-à-dire le pourcentage de séquences reproductibles, a été calculé pour différentes valeurs de p. Ce pourcentage décroît globalement quand p augmente. Comme une valeur élevée de p est recherchée afin de garder des séquences robustes sélectionnées pour beaucoup de jeux de données perturbées, un consensus doit être trouvé. La plus grande valeur de p permettant une reproductibilité acceptable doit donc être choisie. Sur la Figure 3.19, on observe que la décroissance du pourcentage de séquences reproductibles s'accroît soit entre $p = 80\%$ et $p = 90\%$, soit entre

p = 90% et p = 95%. Selon les méthodes et les jeux de données, les valeurs p = 80% ou p = 90% devraient donc être choisies. Afin d'homogénéiser les résultats et d'être sûr de la reproductibilité, la valeur p = 80% a finalement été sélectionnée.

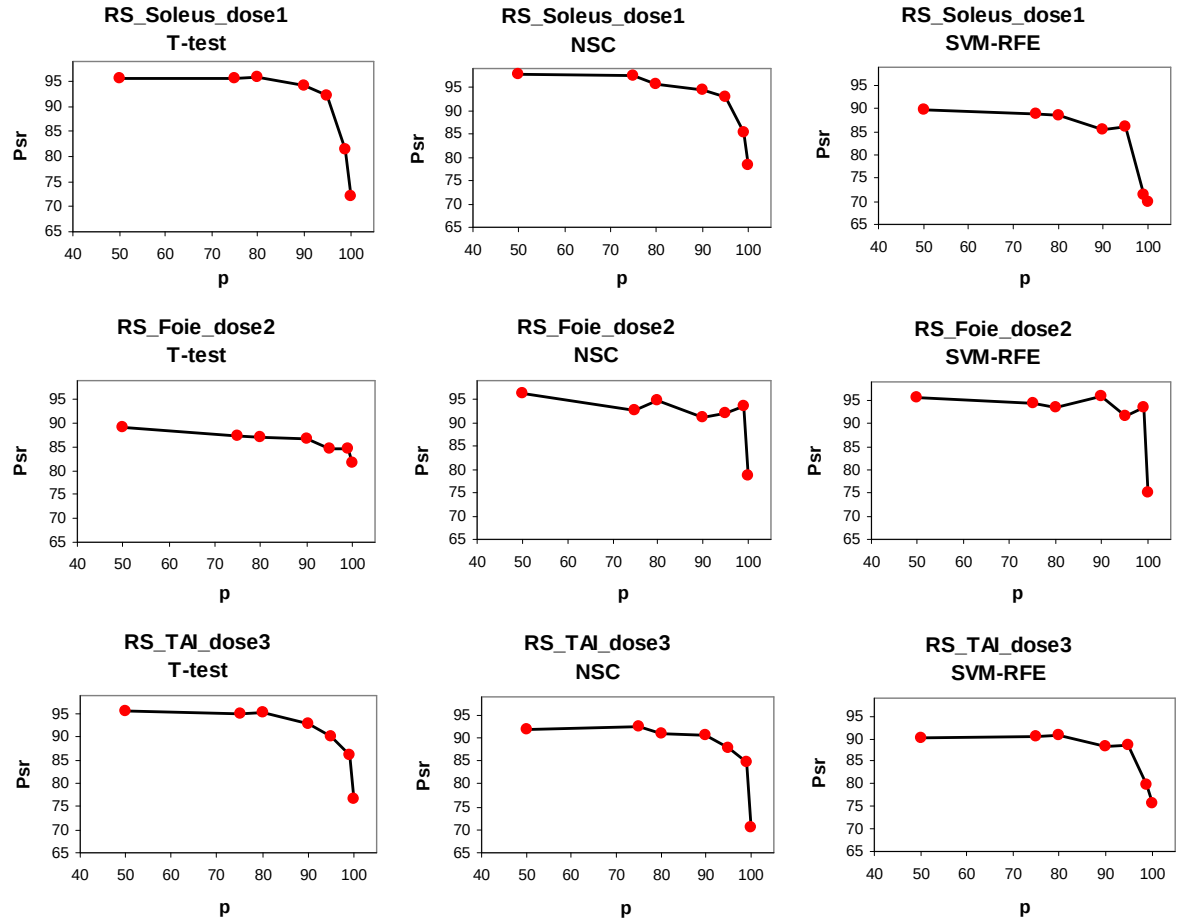

FIG. 3.20 : Reproductibilité à travers différents lancements en fonction de p

Psr est le pourcentage de séquences reproductibles et p est le pourcentage des jeux de données perturbées dans lesquels une séquence doit être trouvée afin d'être sélectionnée. Le pourcentage de séquences reproductibles correspond aux séquences qui ont été trouvées dans les trois lancements de 300 perturbations. Les résultats sont présentés pour NSC, T-test et SVM-RFE dans les jeux de données RS_Soleus_dose1, RS_Foie_dose2 et RS_TAI_dose3.

3.4 Conclusion

La méthodologie MetRob permet d'obtenir une liste des séquences expliquant au mieux les différences entre deux classes de données (ici deux traitements pour une dose et un organe donnés). En résumé, après pré-traitement, les données sont perturbées empiriquement 300 fois. Chaque méthode de sélection de variables (T-test, NSC et SVM-RFE) est appliquée au jeu de données réel et aux jeux de données perturbées afin d'obtenir 301 classements des séquences statistiquement significativement régulées. Une robustesse est ensuite calculée pour toutes les longueurs de listes de séquences de 10 en 10. Ces valeurs permettent le choix d'une longueur de liste de séquences L, maximisant la robustesse pour un nombre minimal de séquences. 301 listes des L meilleures séquences sont donc obtenues pour chaque jeu de données perturbées ou non perturbées. Enfin, les séquences de la liste obtenue sur données réelles sont conservées uniquement si elles sont présentes dans au moins 80% des résultats sur données perturbées. Ces paramètres de MetRob ont été choisis afin d'optimiser robustesse et reproductibilité des listes finales de séquences obtenues avec NSC, T-test et SVM-RFE. MetRob a principalement été appliquée aux neuf jeux de données de puces à ADN présentés précédemment et permettant la comparaison de la rosiglitazone et du SCOMP à trois doses différentes dans trois organes. Ces résultats ainsi que ceux ayant été obtenus sur des données simulées sont présentés dans le Chapitre 4.

Chapitre 4 :

Résultats : Efficacité des méthodes

La méthodologie MetRob a été développée autour de trois méthodes de sélection de variables : T-test, Nearest Shrunken Centroids (NSC) et Support Vector Machine – Recursive Feature Elimination (SVM-RFE). Les performances de ces trois méthodes ont été comparées en termes de robustesse, de pouvoir discriminant des séquences sélectionnées et de pertinence de résultats. Pour ce faire, des tests ont été réalisés sur des données simulées et sur les neuf jeux de données réelles de l'étude PPAR. L'apport de la méthodologie MetRob a également été évalué et, en parallèle, une étude de l'impact sur les résultats du nombre d'animaux considérés a été initiée.

4.1 Génération des données simulées

Les trois méthodes de sélection de variables ont été évaluées sur des jeux de données artificielles : n lames de 4400 séquences, appartenant pour moitié à une classe 0 et pour moitié à une classe 1. Un jeu de données de base de n lames provenant d'une même condition expérimentale a dans un premier temps été simulé avec le logiciel SIMAGE [70]. Des séquences discriminant les deux classes ont ensuite été introduites de manière aléatoire.

4.1.1 Génération des données de base : SIMAGE

4.1.1.2 Principe

SIMAGE est un logiciel disponible en ligne [71] qui permet de simuler des données de puces à ADN en reproduisant les différents facteurs qui influencent une expérience. Il a essentiellement été développé dans le but d'optimiser des protocoles expérimentaux ou de comparer des méthodes d'analyse de données. SIMAGE simule des données de puces cDNA (puces sur lesquelles sont déposées de longues séquences d'ADN complémentaire : cf 1.4.1), différentes des puces à oligonucléotides que nous avons utilisées. Néanmoins, la majorité des biais expérimentaux introduits existent pour les deux types de lames et l'approximation a

donc été considérée comme suffisante. Cette simulation, qui génère des intensités de signaux et des log ratios, permet d'introduire ultérieurement les séquences discriminantes.

SIMAGE met en œuvre plusieurs transformations pour générer des données qui reflètent les différents facteurs influençant une expérience de puce à ADN. Chaque lame est composée N spots, correspondant chacun à une séquence, pour lesquels trois valeurs sont modélisées : log de l'intensité du canal Cy5, log de l'intensité du canal Cy3 et log ratio. Des valeurs de base sont tout d'abord générées pour les différents signaux : chaque séquence reçoit pour le canal k une valeur Gsk tirée dans une loi normale de moyenne μ et d'écart type σ_G (paramètres de la méthode). Cette valeur de base est conservée pour les séquences non régulées. Une proportion de séquences « up-régulées » (plus exprimées dans l'échantillon marqué Cy5) et de séquences « down-régulées » (moins exprimées dans l'échantillon marqué Cy3) est définie par l'utilisateur. Les séquences régulées sont choisies aléatoirement et une valeur fixe μ_D est ajoutée ou soustraite aux log intensités selon le canal et le sens de régulation. Le log ratio, qui est le rapport entre la log intensité Cy5 et la log intensité Cy3, est donc diminué ou augmenté de $2\mu_D$ pour ces séquences.

A ce stade, seule une variation biologique naturelle entre les séquences a été modélisée. Deux séquences identiques, mais ayant des positions différentes sur la lame, ont reçu la même valeur. Une erreur aléatoire est donc ajoutée afin de représenter la variation naturelle liée aux différences dépendant de l'emplacement sur la lame. Elle permet également de prendre en compte des facteurs inconnus et est tirée dans une loi normale centrée en 0 et d'écart type σ_ε (paramètre de la méthode). Seules ces étapes ont été utilisées pour notre simulation. En effet, les autres transformations ajoutent des biais qui ont été corrigés par le logiciel Feature Extraction pour nos données (cf 1.5.1). Parmi ces étapes supplémentaires, un bruit de fond de surface et une déviation liée aux différences entre fluorochromes sont normalement ajoutés au signal.

4.1.1.2 Caractéristiques des données générées

Nous avons généré 18 lames de 4400 séquences, dont 25% sont « up-régulées » et 25% sont « down-régulées ». Les paramètres par défaut de SIMAGE ont été utilisés, mais la plupart des options ont été désactivées (voir paragraphe précédent). Ces paramètres ont été évalués par les auteurs du logiciel à partir de données de puces cDNA. L'utilisation de 4400 séquences correspond à une taille de lame par défaut où chaque séquence est représentée une seule fois sur la lame. Le pourcentage de séquences régulées choisi (50% au total) représente une proportion très élevée pour une lame normale de puces à ADN. Il est cependant raisonnable dans le cadre des données réelles PPAR pour lesquelles les séquences ont été pré-filtrées en fonction de leur régulation pour au moins un des deux traitements. En supposant que les deux produits jouent le même rôle, on peut donc attendre au moins 50% de séquences régulées pour chacun. Enfin, le choix de 18 lames séparées en deux classes est également corrélé aux

données réelles : 6 animaux par traitement pour les jeux de données définis au Chapitre 1 (cf 1.7), mais jusqu'à 9 animaux par traitement disponibles pour le foie.

Les données obtenues présentent une intensité moyenne (moyenne des intensités Cy3 et Cy5) comprise entre 99.746 et 104048.6, alors que celle des données réelles varie entre 9.233 et 376844.93. Comme expliqué précédemment, les signaux simulés ont été calculés à partir d'une même intensité de base pour toutes les séquences, ce qui restreint leur amplitude. Les intensités moyennes devront donc être rééchelonnées avant de pouvoir appliquer la perturbation des données de MetRob qui utilise ces valeurs pour choisir un bruit sur le log ratio. D'autre part, les valeurs de log ratios sont comprises entre -0.717 et 0.765. Cette échelle devra également être prise en compte pour l'ajout de variables discriminantes d'amplitude réaliste. La distribution des log ratios pour une lame de données simulées (Figure 4.1.a) est un peu plus large que celle d'une lame de données réelles (Figure 4.1.b), mais cette approximation sera considérée comme suffisante pour la suite.

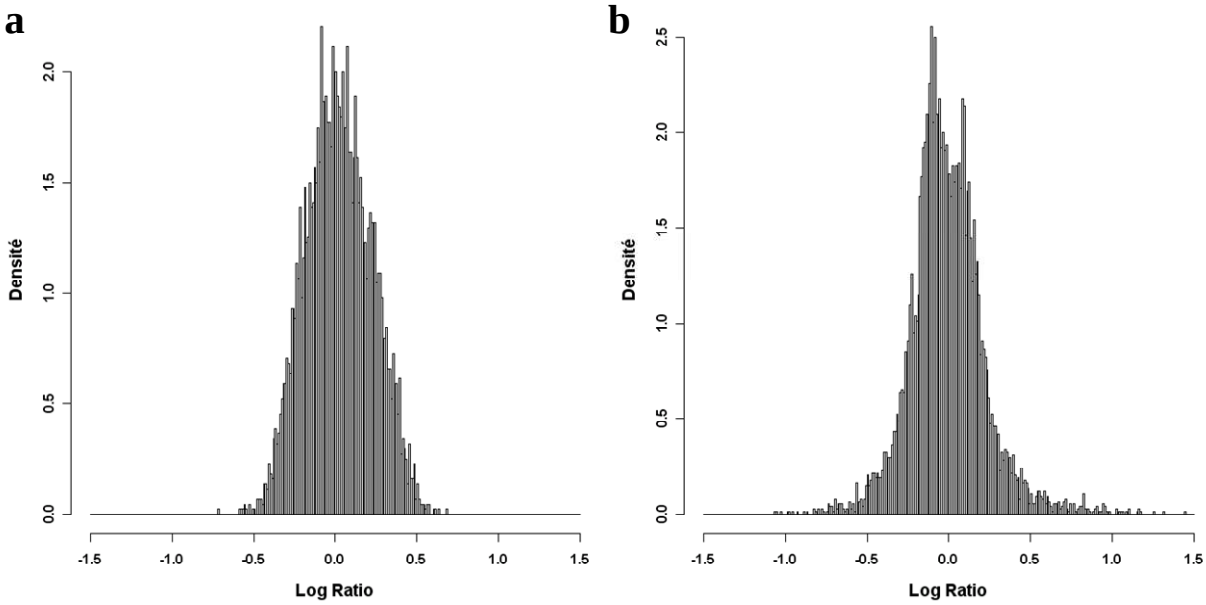

FIG. 4.1 : Histogrammes du log ratio pour une lame de donnée simulées (a) et pour une lame de données réelles (b)
Pour les deux histogrammes, les barres sont de largeur 0.01

4.1.2 Introduction des séquences discriminantes

SIMAGE a permis de générer des données de puces à ADN correspondant à une même condition expérimentale. Comme une répartition en deux classes était recherchée, des variables discriminantes ont été introduites. Les 18 lames initiales ont été séparées en deux groupes de taille égale, puis les valeurs de certaines variables ont été modifiées de manière à

introduire des différences entre les deux classes. Pour cela, les valeurs de log ratio ont été modifiées pour M séquences ($M \in [100, 500, 1000]$) alternativement dans les classes 0 et 1. La différence d introduite entre les log ratios des deux classes a été tirée aléatoirement dans une loi normale centrée en m et d'écart type $\sigma = 0.005$. Deux intervalles ont été choisis pour m afin d'étudier l'impact de l'amplitude de la variation sur sa détection par MetRob : $m \in [0.05, 0.1, 0.15, 0.2, 0.25]$ ou $m \in [0.2, 0.25, 0.3, 0.35, 0.4]$. Pour chaque séquence, m n'a pas une valeur fixe. Il est tiré dans l'intervalle choisi car dans la réalité l'écart entre les classes n'est pas identique pour toutes les séquences. Les intensités Cy3 et Cy5 (respectivement I_g et I_r) ont ensuite été modifiées de la manière suivante : $I_g = I_g \times 10^{-d/2}$ et $I_r = I_r \times 10^{d/2}$, afin de conserver la relation $\text{LogRatio} = \log(\frac{I_r}{I_g})$. Enfin, la moyenne de ces intensités, M_I, a été quadratiquement rééchelonnée pour correspondre à celle obtenue pour les données réelles : $\text{new\_M}_I = 3.481 \times 10^{-5} \times \text{old\_M}_I^{\,2} + 8.887$ (Figure 4.2). Six jeux de données ont donc été obtenus, en ne considérant dans un premier temps que 6 lames par classe (Tableau 4.1).

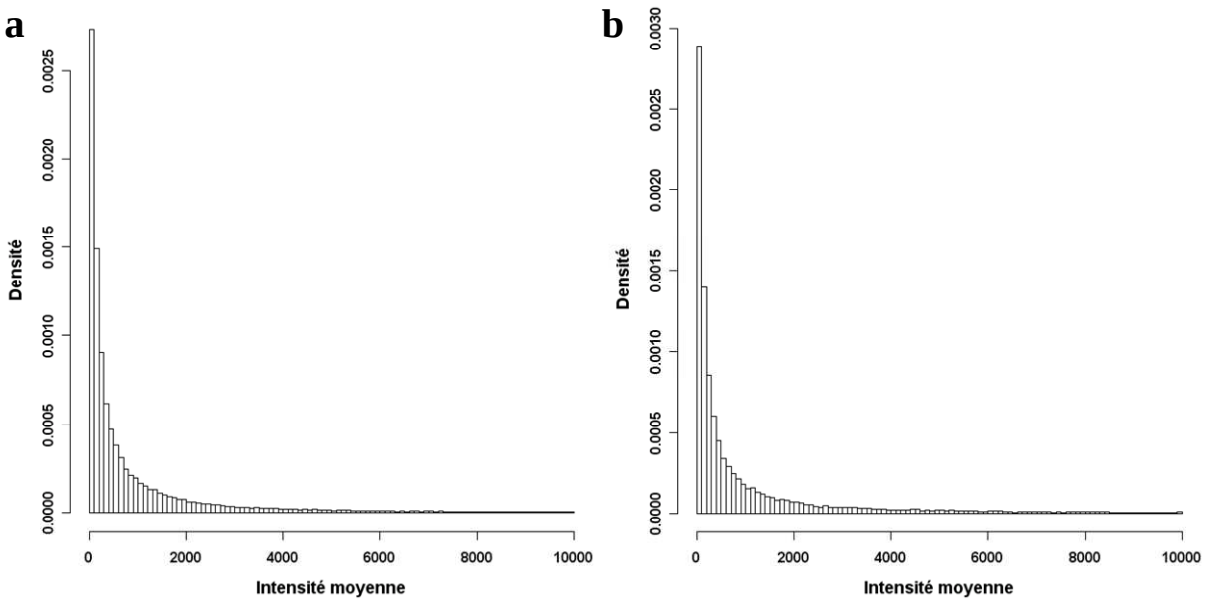

FIG. 4.2 : Histogrammes de la moyenne des intensités Cy3 et Cy5 sur données simulées après rééchelonnement (a) et sur données réelles (b)
Les histogrammes ont été limités aux intensités inférieures à 10^4 pour des raisons de lisibilité. Les barres sont de largeur 100.

σ = 0.005	**m ∈ [0.05, 0.1, 0.15, 0.2, 0.25]**			**m ∈ [0.2, 0.25, 0.3, 0.35, 0.4]**		
	M = 100	**M = 500**	**M = 1000**	**M = 100**	**M = 500**	**M = 1000**
	Sim1_6	Sim2_6	Sim3_6	Sim4_6	Sim5_6	Sim6_6

TAB. 4.1 : Nomenclature des jeux de données simulées avec 6 lames par classe

Les histogrammes des différences de log ratio moyen entre les classes ont été représentés sur la Figure 4.3 pour les jeux de données Sim2_6 et Sim5_6. Les histogrammes en rouge et noir correspondent respectivement aux séquences discriminantes et non discriminantes. Pour des différences de log ratio moyen faibles (Figure 4.3.a), les deux histogrammes ont une zone de recouvrement importante. En revanche, ils sont presque distincts (Figure 4.3.b) dans le cas de différences plus importantes. Les deux types de séquences sont donc à priori plus difficiles à distinguer dans le premier cas. Sur données réelles, l'histogramme des différences de log ratio moyen entre les classes a l'allure représentée sur la Figure 4.4. On ne distingue pas de pic pouvant être lié à des séquences discriminantes, qui seraient alors facilement identifiables. Les jeux de données simulées introduisant peu de différences entre les log ratios sont donc les plus proches de la réalité.

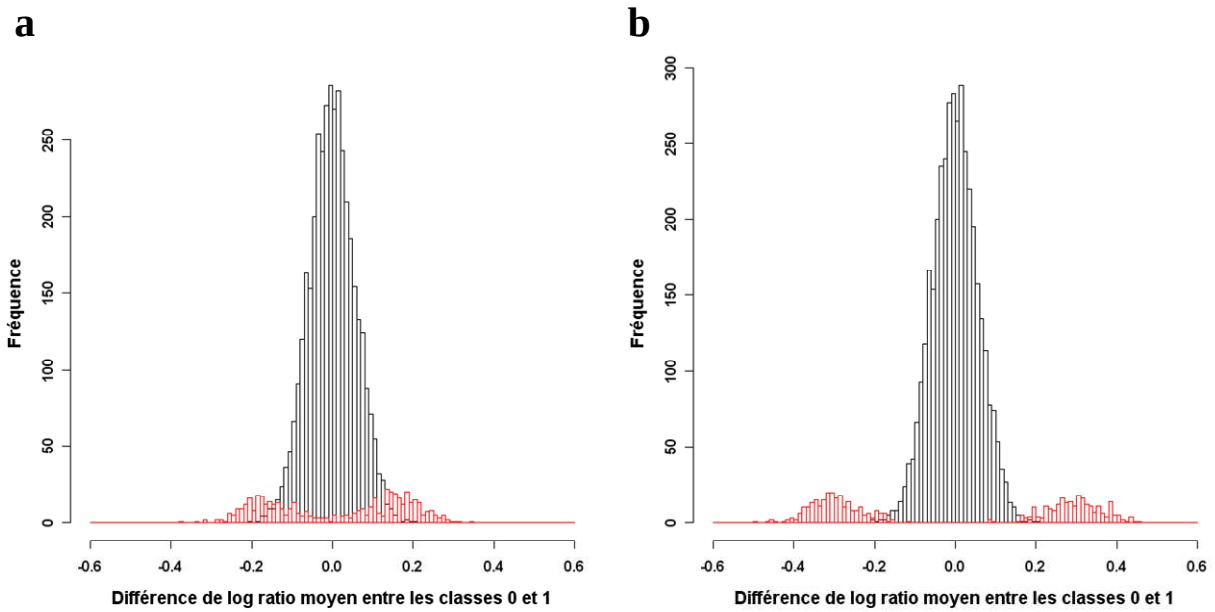

FIG. 4.3 : Histogrammes des différences de log ratio moyen entre la classe 0 et la classe 1 pour les jeux de données Sim2_6 (a) et Sim5_6 (b)
Les variables non discriminantes sont représentées en noir et les variables discriminantes en rouge. Pour les deux histogrammes, les barres sont de largeur 0.01.

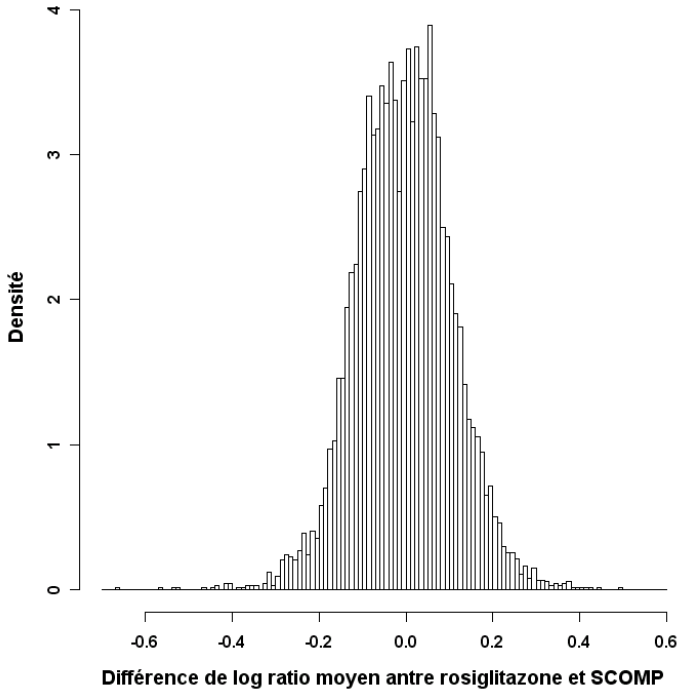

FIG. 4.4 : Histogramme des différences de log ratio moyen entre la rosiglitazone et le SCOMP pour le jeu de données RS_Soleus_dose1
Les barres sont de largeur 0.01.

4.2 Résultats sur données simulées

La méthodologie MetRob, couplée au T-test, à NSC ou à la SVM-RFE, a été testée sur les six jeux de données simulées présentés précédemment. Les trois méthodes de sélection de variables ont été évaluées en fonction de leur robustesse, mais aussi en fonction du pouvoir discriminant et de la pertinence des variables sélectionnées. Les caractéristiques de ces variables ont ensuite été étudiées, ainsi que l'impact sur les résultats du nombre d'observations prises en compte.

4.2.1 Robustesse des méthodes de sélection de variables

Dans MetRob, les méthodes de sélection de variables sont utilisées pour classer les séquences, puis la robustesse est calculée pour toutes les longueurs de listes de séquences de 10 en 10. La robustesse est le pourcentage moyen de séquences communes entre une liste de séquences obtenue sur données non perturbées et une liste de séquences obtenue sur données perturbées. On cherche à étudier la robustesse des trois méthodes testées (T-test, NSC et SVM-RFE) pour un nombre de séquences donné.

	Sim1_6			Sim2_6			Sim3_6		
	T-test	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**
10	63.5	60.57	40.4	41.57	44.83	23.7	49.77	59.73	40.03
100	70.69	70.18	67.78	73.42	73.75	69.42	75.69	77.22	72.97
500	73.11	73.3	72.64	80.8	81.11	79.73	83.38	83.45	82.24
1000	77.92	77.96	75.45	81.58	81.58	80.45	84.6	84.63	84.46
2000	82.33	82.32	81.05	84.27	84.27	83.28	86.68	86.66	85.73
3000	86.99	86.99	79.63	87.95	87.94	80.76	88.99	88.99	82.38
4000	93.54	93.54	92.16	93.81	93.81	92.35	94.07	94.07	92.74
4400	100	100	100	100	100	100	100	100	100

	Sim4_6			Sim5_6			Sim6_6		
	T-test	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**
10	80.2	75.6	58.3	65.77	66.43	41.87	59.27	62.97	46.87
100	82.98	84.33	81.45	77.48	79.13	73.63	70.42	72.16	69.45
500	74.48	74.65	74.21	91.28	91.26	91.32	86.02	85.58	85.42
1000	78.13	78.18	75.92	85.32	85.37	84.82	94.37	94.54	94.51
2000	82.41	82.4	81.07	85.02	85.01	83.81	88.8	88.8	87.92
3000	87.02	87.01	79.6	88.16	88.16	81.15	89.61	89.61	83.37
4000	93.53	93.53	92.17	93.77	93.78	92.42	94.3	94.3	92.86
4400	100	100	100	100	100	100	100	100	100

TAB. 4.2 : Robustesse des trois méthodes de sélection de variables pour différentes longueurs de listes de séquences

Vert : méthode la plus robuste. Jaune : méthode intermédiaire. Orange : méthode la moins robuste. La ligne encadrée correspond à 3000 séquences et est expliquée dans le corps du manuscrit.

Cette robustesse croît globalement avec le nombre de séquences jusqu'à atteindre logiquement 100% quand toutes les séquences sont prises en compte (Tableau 4.2). Quelque soit le jeu de données, la méthode SVM-RFE obtient la robustesse la plus faible pour presque toutes les longueurs de listes de séquences. On peut noter que l'écart entre la SVM-RFE et les deux autres méthodes augmente pour 3000 séquences (lignes encadrées du Tableau 4.2). Cette observation est en fait un artefact lié au choix des paramètres de la SVM-RFE : pour des raisons de complexité de calcul, seule la première moitié des séquences est classée pertinemment (cf 3.3.3.3). Les méthodes T-test et NSC ont par ailleurs des résultats de robustesse assez proches. Pour 10 séquences, des différences plus importantes peuvent apparaître, mais la méthode la plus robuste dépend du jeu de données : le T-test est plus robuste pour Sim1_6 et Sim4_6 et NSC est plus robuste pour Sim2_6, Sim3_6, Sim5_6 et Sim6_6. On peut donc globalement conclure que la méthode de SVM-RFE est moins robuste que les deux autres, mais qu'il est difficile de différencier le T-test et NSC sur les jeux de données simulées.

4.2.2 Pertinence des listes de séquences sélectionnées

La qualité de la sélection des variables a ensuite été évaluée. Pour cela, les quantités suivantes ont été définies (Tableau 4.3) :

- Vrais Positifs (VP) : séquences sélectionnées et réellement discriminantes
- Faux Positifs (FP) : séquences sélectionnées mais non discriminantes
- Vrais Négatifs (VN) : séquences non sélectionnées et non discriminantes
- Faux Négatifs (FN) : séquences non sélectionnées mais discriminantes

	Séquences réellement discriminantes	Séquences réellement non discriminantes
Séquences sélectionnées	**VP**	**FP**
Séquences non sélectionnées	**FN**	**VN**

TAB. 4.3 : Evaluation des résultats des méthodes de sélection de variables
VP = vrais positifs, FP = faux positifs, VN = vrais négatifs, FN = faux négatifs

Ces valeurs permettent de calculer différents scores qui quantifient la qualité des méthodes de sélection de variables (Figure 4.5). La capacité d'une méthode à détecter les variables discriminantes est évaluée par la sensibilité, proportion de séquences discriminantes trouvées parmi toutes les séquences réellement discriminantes. La spécificité est son équivalent pour les séquences non discriminantes. La confiance accordée aux prédictions est donnée par la valeur prédictive positive (respectivement négative), qui correspond à la proportion de séquences réellement discriminantes (respectivement non discriminantes) parmi les séquences sélectionnées comme telles.

$$\mathrm{Se} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \qquad \mathrm{Sp} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}}$$

$$\mathrm{VPP} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \qquad \mathrm{VPN} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FN}}$$

FIG. 4.5 : Scores quantifiant la qualité des méthodes de sélection de variables
Se = sensibilité, Sp = spécificité, VPP = valeur prédictive positive, VPN = valeur prédictive négative

Les résultats de MetRob concernant ces scores ont donc été exploités pour les trois méthodes de sélection de variables (Tableau 4.4). Les spécificités sont excellentes ($Sp > 0.99$), essentiellement car le nombre de faux positifs est négligeable par rapport au total des séquences non discriminantes. De même, les valeurs prédictives négatives sont globalement bonnes. Elles diminuent pour Sim3_6 et Sim6_6 car le nombre de vrais négatifs diminue alors que le nombre de faux négatifs augmente. En revanche, les sensibilités restent faibles dans tous les cas ($Se < 0.6$). Beaucoup de séquences discriminantes ne sont donc pas détectées. Les résultats sont meilleurs pour les jeux de données Sim4_6, Sim5_6 et Sim6_6, qui correspondent à des écarts de log ratio plus importants entre les classes.

		Se	Sp	VPP	VPN
Sim1_6	**T-test**	0.32	0.991	0.457	0.984
	NSC	0.32	0.991	0.464	0.984
	SVM-RFE	0.25	0.995	0.543	0.983
Sim2_6	**T-test**	0.264	0.993	0.835	0.913
	NSC	0.278	0.993	0.842	0.915
	SVM-RFE	0.228	0.996	0.891	0.91
Sim3_6	**T-test**	0.087	0.999	0.978	0.788
	NSC	0.07	0.999	0.972	0.785
	SVM-RFE	0.128	0.999	0.977	0.796
Sim4_6	**T-test**	0.44	1	0.978	0.987
	NSC	0.31	1	1	0.984
	SVM-RFE	0.54	1	0.982	0.989
Sim5_6	**T-test**	0.55	0.999	0.993	0.945
	NSC	0.594	0.999	0.993	0.95
	SVM-RFE	0.56	1	0.996	0.947
Sim6_6	**T-test**	0.434	1	1	0.855
	NSC	0.477	1	0.998	0.867
	SVM-RFE	0.465	1	1	0.864

TAB. 4.4 : Pertinence des listes de séquences obtenues avec MetRob pour les trois méthodes de sélection de variables et les six jeux de données simulées
Se : sensibilité, Sp : spécificité, VPP : valeur prédictive positive, VPN : valeur prédictive négative.

Les résultats obtenus en termes de valeurs prédictives positives (VPP) sont meilleurs : les séquences sélectionnées sont en grande majorité réellement discriminantes. Seul le jeu de données Sim1_6 présente de moins bonnes valeurs. L'intérêt essentiel de cette méthodologie est donc l'obtention d'une bonne VPP, c'est-à-dire de résultats certes incomplets, mais fiables. Les situations où il existe très peu de différences (peu de variables et peu d'écart entre

les classes) restent néanmoins difficiles à traiter. Les différences de VPP entre les trois méthodes de sélection de variables sont en général minimes. La SVM-RFE obtient néanmoins un résultat légèrement meilleur dans le cas cité précédemment où il y a peu de différences entre les classes.

La pertinence des listes de séquences sélectionnées par MetRob a été évaluée dans l'absolu. On peut néanmoins se demander ce que la méthodologie apporte par rapport aux méthodes de sélection de variables utilisées seules. La pertinence des listes de séquences en termes de sensibilité et de VPP est présentée dans le Tableau 4.5 pour la méthode NSC. Les conclusions obtenues avec les autres méthodes sont similaires et ne sont donc pas présentées ici. Les séquences ont été sélectionnées de deux manières différentes :

- MetRob couplée à NSC : NSCRob
- Méthode NSC utilisée seule : les M meilleures séquences ont été sélectionnées (M : nombre réel de variables discriminantes)

		Sim1_6	**Sim2_6**	**Sim3_6**	**Sim4_6**	**Sim5_6**	**Sim6_6**
Se	**NSCRob**	0.32	0.278	0.07	0.31	0.594	0.477
	NSCSeule	0.42	0.622	0.687	0.83	0.894	0.944
VPP	**NSCRob**	0.464	0.842	0.972	1	0.993	0.999
	NSCSeule	0.42	0.622	0.687	0.83	0.894	0.944

TAB. 4.5 : Comparaison de la pertinence des séquences trouvées avec et sans l'utilisation de MetRob, dans le cas de Nearest Shrunken Centroids
Se : sensibilité, VPP : valeur prédictive positive

Les sensibilités obtenues avec MetRob sont moins bonnes que pour la méthode NSC utilisée seule. En effet, appliquer MetRob réduit considérablement le nombre de séquences sélectionnées et, en parallèle, le nombre de séquences réellement discriminantes sélectionnées. A titre d'exemple, au lieu des 1000 séquences attendues, NSCRob en sélectionne 72 pour Sim3_6 et 478 pour Sim6_6. En revanche, les valeurs prédictives positives sont améliorées par l'utilisation de MetRob. Ceci confirme que l'intérêt principal de la méthodologie réside dans l'obtention de séquences fiables. De plus, l'apport de MetRob en termes de VPP est plus important pour les jeux de données Sim2_6 et Sim3_6. Cette nouvelle méthodologie est donc particulièrement utile pour des distributions de séquences discriminantes et de séquences non discriminantes qui se recoupent, ce qui correspond le plus à des situations réelles.

4.2.3 Etude des listes de séquences sélectionnées

4.2.3.1 Comparaison des trois méthodes de sélection de variables

Les séquences sélectionnées par la méthodologie MetRob ont ensuite été étudiées plus précisément. Dans un premier temps, les listes obtenues avec les trois méthodes de sélection de variables ont été comparées. Pour cela, l'ensemble S des séquences trouvées, toutes méthodes confondues, a été considéré pour chaque jeu de données. Les pourcentages de séquences sélectionnées par les trois méthodes, par deux méthodes ou par une seule des méthodes ont été calculés par rapport à S (Figure 4.6).

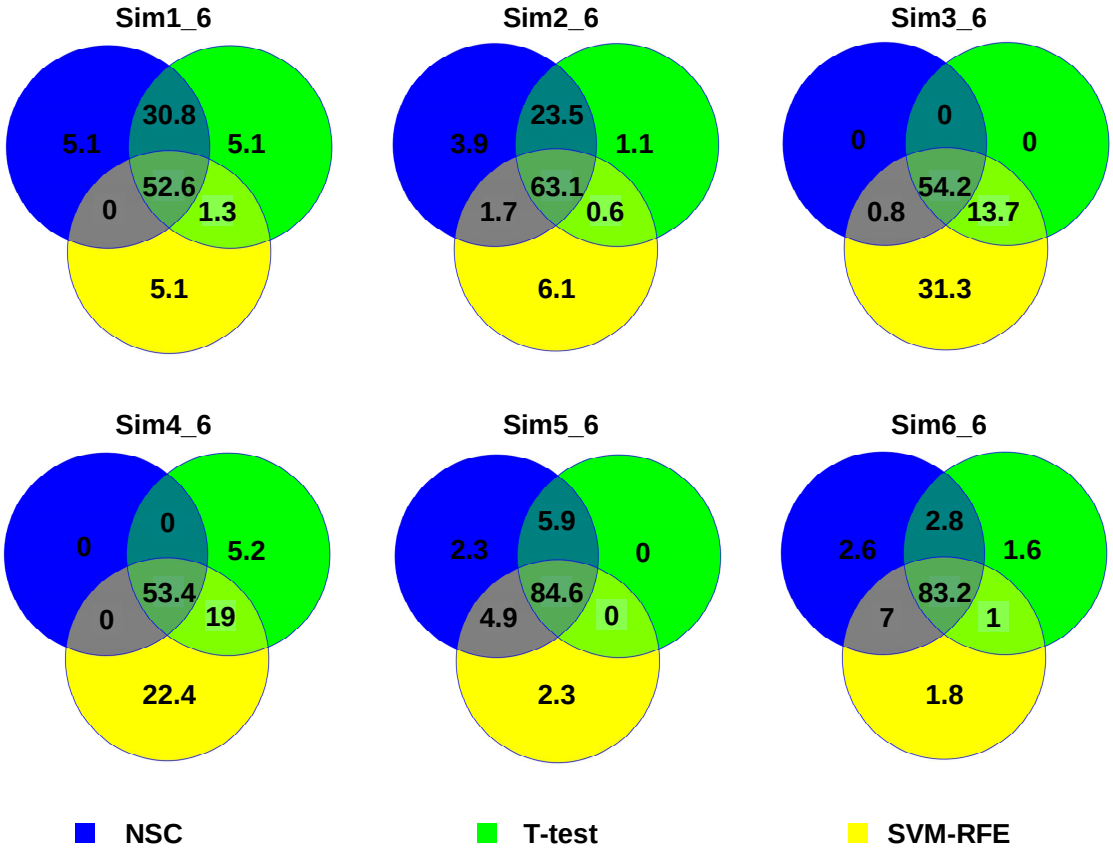

FIG. 4.6 : Comparaison des listes de séquences obtenues avec les trois méthodes de sélection de variables pour tous les jeux de données simulées
Le pourcentage de chaque groupe de séquences a été calculé par rapport à l'ensemble des séquences trouvées pour un même jeu de données toutes méthodes confondues.

Pour les jeux de données Sim1_6 et Sim2_6, la configuration est sensiblement la même : la majorité des séquences sont communes au T-test et à NSC et plus de la moitié des séquences ont été sélectionnés par les trois méthodes. Pour les jeux de données Sim3_6 et Sim4_6, la méthode de SVM-RFE a sélectionné entre 23.6% et 31.3% de séquences qui lui sont propres et la majorité des séquences obtenues avec NSC ont également été trouvées avec les autres méthodes. Enfin, sur les deux derniers jeux de données, Sim5_6 et Sim6_6, plus de 80% des séquences ont été sélectionnées par les trois méthodes de sélection de variables. L'impact du choix d'une méthode dépend donc des données considérées. Dans des situations de discrimination facile (beaucoup de séquences et beaucoup d'écart entre les classes), les trois méthodes sont à peu près équivalentes. En revanche, dans les autres cas, SVM-RFE se distingue du T-test et de NSC en sélectionnant soit plus de séquences, soit moins de séquences. Cette méthode se positionne donc à part des deux autres.

4.2.3.2 Différences de log ratio entre les classes

Dans un second temps, nous avons voulu vérifier si les séquences présentant le plus de différences entre les classes avaient bien été choisies. Dans cette optique, pour chaque méthode de sélection de variables, la liste des séquences réellement discriminantes sélectionnées par MetRob (L_m) et la liste de toutes les séquences réellement discriminantes (L) ont été classées par écart de log ratio moyen entre les classes décroissant. Ces listes ont ensuite été comparées à l'aide du critère C suivant : $C = \frac{\sum_{n=1}^{N} A_n \times e^{-\alpha n}}{\sum_{n=1}^{N_m} n \times e^{-\alpha n} + \sum_{n=N_m+1}^{N} N_m \times e^{-\alpha n}}$, où N est le nombre de séquences réellement discriminantes, N_m est le nombre de séquences réellement discriminantes sélectionnées par MetRob et A_n est le nombre de séquences communes aux deux listes parmi les n premières [72]. Alpha a été choisi de manière à négliger les séquences classées après N_m ($e^{-\alpha N_m} < 10^{-2}$). Le numérateur représente l'adéquation du classement des deux listes et le dénominateur est le score idéal qui serait obtenu si les deux listes étaient identiquement classées.

Une valeur de C élevée indique que les séquences choisies par MetRob sont en priorité celles qui présentent les différences de log ratio moyen les plus importantes entre les deux classes. A titre d'exemple, pour $N = 500$ et $N_m = 150$, si les 75 premières séquences de L_m sont identiques aux 75 premières séquences de L et que les autres séquences de L_m sont les 75 dernières séquences de L qui n'interviennent donc pas dans le score, C vaut 0.909 (Figure 4.7.a). Si les 75 premières et les 75 dernières séquences de L sont classées alternativement dans L_m, alors C vaut 0.508 (Figure 4.7.b).

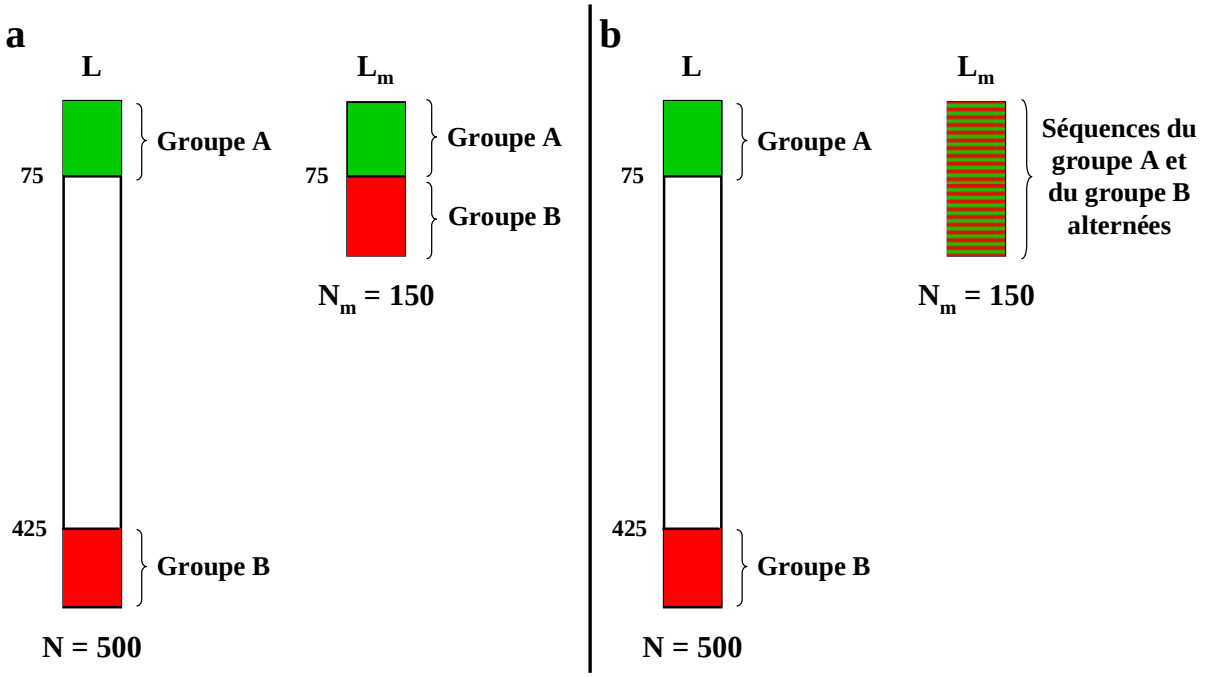

FIG. 4.7 : Exemples de situations pour le calcul de C
Groupe A : 75 premières séquences de L. Groupe B : 75 dernières séquences de L.
(a) : C vaut 0.909. (b) : C vaut 0.508.

Le tableau 4.6 présente les valeurs de C obtenues avec les trois méthodes de sélection de variables pour tous les jeux de données simulées. Globalement, C est plus faible pour les jeux de données correspondant à des écarts plus faibles entre les classes (Sim1_6, Sim2_6 et Sim3_6) et il diminue quand le nombre de variables discriminantes augmente. Excepté pour le jeu de données Sim3_6, pour lequel il est plus faible, le pourcentage de séquences communes entre les N_m séquences sélectionnées et les N_m séquences les plus différentes entre les classes est aux alentours de 50%. Cette situation est proche des exemples de valeurs de C citées dans le paragraphe précédent et permet d'avoir un point de comparaison. Le critère C est, sauf pour Sim3_6, supérieur à 0.6 et peut atteindre 0.825, ce qui suggère effectivement que les séquences ayant des différences de log ratios moyens plus importantes entre les classes ont été sélectionnées. En revanche, pour Sim3_6, cette affirmation n'est pas vérifiée. On peut donc supposer que, pour des écarts faibles et beaucoup de séquences discriminantes, ce ne sont plus nécessairement les séquences présentant le plus grand écart de log ratio moyen entre les classes qui sont sélectionnées par MetRob.

C	T-test	NSC	SVM-RFE
Sim1_6	0.71	0.662	0.734
Sim2_6	0.665	0.71	0.603
Sim3_6	0.472	0.466	0.586
Sim4_6	0.825	0.803	0.783
Sim5_6	0.766	0.783	0.773
Sim6_6	0.665	0.667	0.646

TAB. 4.6 : Valeurs du critère C pour les trois méthodes de sélection de variables et les six jeux de données simulées

4.2.4 Pouvoir discriminant des listes de séquences

Les caractéristiques des listes de séquences obtenues avec MetRob ont été étudiées, mais leur capacité à classer correctement les observations reste encore à être validée. Les séquences sélectionnées par MetRob (couplée au T-test, à NSC ou à SVM-RFE) ont donc été utilisées pour classer les données. La classification a été réalisée au moyen d'un SVM linéaire et l'erreur obtenue par Leave-One-Out-Cross-Validation (LOOCV) a été calculée. Les résultats ont été comparés à la qualité d'une discrimination utilisant les 4400 séquences initiales (Tableau 4.7).

	Toutes les séquences	Séquences MetRob		
		T-test	NSC	SVM-RFE
Sim1_6	0.417	0	0	0
Sim2_6	0	0	0	0
Sim3_6	0	0	0	0
Sim4_6	0.167	0	0	0
Sim5_6	0	0	0	0
Sim6_6	0	0	0	0

TAB. 4.7 : Erreurs par LOOCV

Les erreurs par LOOCV obtenues avec les séquences de MetRob sont nulles quels que soient le jeu de données et la méthode de sélection de variable utilisée. De même, excepté pour les jeux de données Sim1 et Sim4 qui comportent peu de séquences discriminantes, l'erreur de discrimination est nulle avec toutes les séquences. Le pouvoir discriminant des séquences obtenues par MetRob est ainsi validé, mais ne démontre pas de réelle supériorité par rapport à l'ensemble des séquences. On retrouve essentiellement ce qui avait été énoncé au Chapitre 2 (cf 2.3) : le choix du nombre de séquences à conserver ne peut être effectué en se basant sur une erreur de classification nulle.

4.2.5 Impact du nombre d'observations

Les jeux de données simulées étudiés jusqu'à présent comportent 6 observations par classe, soit n = 12 lames, de manière à se rapprocher de la situation des données réelles. Plusieurs questions peuvent alors se poser :

- Comment se comportent les résultats si n augmente ou diminue ?
- n est-il suffisant ?
- n est-il optimal ?

Afin de répondre au moins en partie à ces questions, des tests ont été réalisés en considérant 3 observations par classe (n = 6 : Sim1_3, Sim2_3, ..., Sim6_3) puis 9 observations par classe (n = 18 : Sim1_9, Sim2_9, ..., Sim6_9). Les sensibilités et les valeurs prédictives positives (VPP) ont été comparées pour MetRob associée à la méthode NSC (Tableau 4.8).

La sensibilité augmente avec le nombre d'observations pour Sim1, Sim4 et Sim5. En revanche, elle est maximale avec n = 12 pour Sim2 et Sim6 et minimale avec n = 12 pour Sim3. Ces résultats sont donc difficiles à interpréter et doivent être pris avec précaution. En effet, il aurait fallu appliquer MetRob à d'autres jeux de données ayant les mêmes propriétés (même nombre de variables discriminantes et même écart entre les classes) afin d'obtenir des conclusions plus significatives. Ces tests n'ont néanmoins pas pu être effectués par manque de temps. Excepté pour Sim5, la VPP augmente également avec le nombre d'observations. Néanmoins, l'écart est minime entre n = 12 et n = 18, sauf pour Sim1 et Sim2. Considérer 6 observations par classe (n = 12) semble donc déjà fournir de bons résultats en termes de VPP s'il y a beaucoup de différences entre les classes ou beaucoup de variables discriminantes. Afin de répondre plus précisément aux questions posées précédemment, il serait intéressant de tester des valeurs de n intermédiaires (n = 8, 10, 14, 16) et des valeurs plus élevés que n = 18.

		Sim1	**Sim2**	**Sim3**	**Sim4**	**Sim5**	**Sim6**
Se	**n=6**	0.18	0.154	0.132	0.12	0.304	0.232
	n=12	0.32	0.278	0.07	0.31	0.594	0.477
	n=18	0.37	0.206	0.256	0.77	0.838	0.456
VPP	**n=6**	0.121	0.554	0.742	0.857	0.854	0.955
	n=12	0.464	0.842	0.972	1	0.993	0.998
	n=18	0.607	0.981	0.977	0.977	0.998	1

TAB. 4.8 : Comparaison de la pertinence des séquences sélectionnées par MetRob associée à NSC en fonction du nombre d'observations considérées
Se : sensibilité, VPP : valeur prédictive positive.

D'autre part, les listes de séquences obtenues pour n=6, n=12 et n=18, ont été comparées pour les jeux de données Sim2 et Sim5 et la méthode NSC. L'ensemble S des séquences trouvées, toutes valeurs de n confondues, a été considéré pour chaque jeu de données. Les pourcentages de séquences sélectionnées pour les trois valeurs de n, pour deux valeurs de n ou pour une seule valeur de n ont été calculés par rapport à S (Figure 4.8). Pour le jeu de données Sim5, les séquences sélectionnées avec $n = 12$ sont en grande majorité incluses dans celles sélectionnées pour $n = 18$. Ces résultats sont cohérents avec l'augmentation de sensibilité sans réelle perte de VPP constatée précédemment. Dans ce cas, considérer $n = 18$ permet de trouver plus de séquences discriminantes, mais les résultats obtenus pour $n = 12$ sont déjà satisfaisants en termes de VPP. Pour le jeu de données Sim2, la situation est inversée : les séquences sélectionnées pour $n = 18$ sont majoritairement incluses dans celles sélectionnées pour $n = 12$. De même, ces résultats sont cohérents avec l'obtention d'une meilleure VPP pour $n = 18$ et d'une meilleure sensibilité pour $n = 12$. Dans cette situation, considérer 9 observations par classe permet d'améliorer la VPP. Le comportement des résultats quand n croît semble donc dépendre de la nature des données considérées.

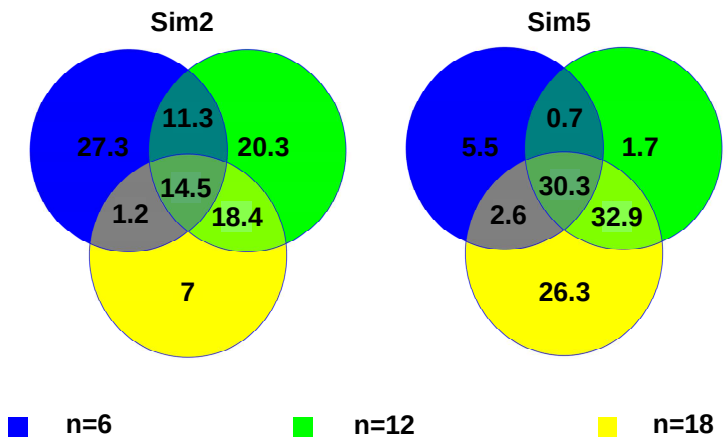

FIG. 4.8 : Comparaison des listes de séquences obtenues avec différents nombres d'observations
Le pourcentage de chaque groupe de séquences a été calculé par rapport à l'ensemble des séquences trouvées tous nombres d'observations confondus. Les résultats sont présentés pour Sim2 et Sim5 avec la méthode NSC.

4.2.6 Conclusions

Les résultats des tests sur données simulées ont permis d'aboutir à plusieurs conclusions qui sont récapitulées ci-dessous :

- La méthode de SVM-RFE fournit des résultats moins robustes que le T-test et la NSC pour un nombre de séquences donné.

- La méthodologie MetRob fournit d'excellentes VPP et est essentiellement utile pour des jeux de données où les distributions d'écart de log ratio moyen entre les classes des séquences non discriminantes et des séquences discriminantes se chevauchent.
- La méthode de SVM-RFE se démarque des deux autres en termes de séquences sélectionnées. Dans la majorité des cas, les séquences ayant le plus fort écart de log ratio moyen entre les classes ont été sélectionnées.
- Une bonne qualité de discrimination est obtenue, même en considérant toutes les séquences. Le choix d'un nombre de séquences optimal ne peut donc pas se restreindre à celui qui minimise l'erreur de classification.
- Le choix de n = 12 semble suffisant pour obtenir une bonne VPP dans les cas où il y a beaucoup de différences entre les classes ou beaucoup de variables discriminantes. Néanmoins, des tests supplémentaires sont nécessaires pour conclure à propos d'un nombre optimal d'observations en fonction du type de données.

4.3 Résultats sur données réelles 4*44k

La méthodologie MetRob a également été appliquée aux données réelles des expériences PPAR : données de comparaison entre la rosiglitazone et le SCOMP présentées dans le Chapitre 1 (cf 1.6.1). Les deux agonistes PPAR ont été comparés à trois doses (28µmol/kg, 84µmol/kg et 280µmol/kg) dans trois organes (muscle squelettique, foie et tissu adipeux inguinal). Chaque jeu de données comporte 6 animaux par traitement. Les séquences ont été pré-filtrées en fonction de leur significativité d'expression par rapport aux animaux contrôles db+ (cf 3.1.1). Le nombre de séquences statistiquement significativement régulées (séquences SSR) obtenu pour chaque cas est donné dans le Tableau 4.9.

	Muscle squelettique	**Foie**	**Tissu adipeux inguinal**
28 µmol/kg	7401	8430	10255
84 µmol/kg	7388	8227	13184
280 µmol/kg	8209	10770	17622

TAB. 4.9 : Nombre de séquences SSR pour chaque dose et chaque organe

Les trois méthodes de sélection de variables ont été comparées sur ces données en termes de robustesse et de listes de séquences sélectionnées. L'impact du nombre d'observations a ensuite été étudié pour le foie, en considérant 6 ou 9 animaux par classe.

4.3.1 Robustesse des méthodes de sélection de variables

De même que pour les jeux de données simulées, les robustesses des trois méthodes de sélection de variables (T-test, NSC et SVM-RFE) ont été comparées pour un nombre de séquences donné (Tableau 4.10).

	RS_Soleus_dose1			RS_Soleus_dose2			RS_Soleus_dose3		
	T-test	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**
10	25.33	36.27	14.77	31.5	30.2	27.17	30.67	36.03	19.37
100	46.87	54.06	49.52	45.99	57.26	50.65	52.44	58.63	48.27
500	71.28	72.64	68.82	65.06	68.92	68.08	70.83	73.08	65.31
2000	88.02	88.27	84.15	85.12	85.43	81.59	85.54	86.51	82.07
5000	94.32	94.3	85.47	94.07	94.1	84.55	95.34	95.3	86.9
7000	96.99	97	95.48	97	96.99	95.44	97.58	97.58	91.45
8000	NA	NA	NA	NA	NA	NA	98.63	98.63	97.65

	RS_Foie_dose1			RS_Foie_dose2			RS_Foie_dose2		
	T-test	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**
10	55.2	60	38.53	44.9	51.9	26.83	45.03	55.77	30.63
100	60.56	62.95	54.47	64.91	66.21	59.21	56	64.09	51.64
500	70.14	71.42	64.86	72.29	73.5	69.18	71.42	73.86	65.66
2000	80.45	80.68	73.86	80.78	80.82	77.5	82.81	83.26	76.6
5000	88.41	88.38	75.37	87.66	87.66	77.57	90.98	91.02	85.24
8000	96.02	96.01	95.21	97.57	97.56	97.34	94.68	94.7	84.3
10000	NA	NA	NA	NA	NA	NA	96.67	96.66	93.86

	RS_TAI_dose1			RS_TAI_dose2			RS_TAI_dose3		
	T-test	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**	**T-test**	**NSC**	**SVM-RFE**
10	74.63	75.53	36.93	41.3	39.4	40.1	33.87	35.93	23.1
100	59.83	60.36	53.69	62.3	64.75	58.14	46.71	47.87	49.33
500	64.86	65.43	61.27	70.32	71.39	64.97	60.2	61.21	62.91
2000	73.22	73.38	69.3	79.53	79.49	73.88	71.25	71.25	72.11
5000	81.19	81.2	77.22	86.58	86.55	80.63	80.13	80.18	77.3
10000	97.79	97.79	97.57	92.89	92.9	83.16	88.44	88.45	76.54
13000	NA	NA	NA	98.77	98.76	98.64	91.55	91.55	81.07
17000	NA	NA	NA	NA	NA	NA	97.21	97.21	94.64

TAB. 4.10 : Robustesse des trois méthodes de sélection de variables pour différentes longueurs de listes de séquences

Vert : méthode la plus robuste. Jaune : méthode intermédiaire. Orange : méthode la moins robuste. Les neuf jeux de données PPAR sont représentés.

Les résultats sont similaires à ceux observées sur données simulées. La robustesse croît globalement avec le nombre de séquences jusqu'à atteindre 100% pour toutes les séquences et la méthode de SVM-RFE présente la robustesse la plus faible dans la majorité des cas. Néanmoins, pour un nombre de séquences supérieur à la moitié des séquences SSR, les mauvais résultats de la SVM-RFE peuvent être expliqués par une absence de classement pertinent après ce seuil. D'autre part, pour des nombres de séquences importants, le T-test et NSC sont presque équivalents. En revanche, pour des nombres de séquences plus faibles, qui sont à priori ceux qui nous intéressent, la méthode NSC est globalement plus robuste que le T-test.

4.3.2 Etude des listes de séquences sélectionnées

4.3.2.1 Comparaison des trois méthodes de sélection de variables

Les séquences sélectionnées par la méthodologie MetRob ont ensuite été étudiées plus précisément en comparant les listes obtenues avec les trois méthodes de sélection de variables. La comparaison est présentée pour le foie sur la Figure 4.9. Le T-test et la méthode de NSC ont des résultats très proches sur les jeux de données RS_Foie_dose1 et RS_Foie_dose2. Même si le T-test est un peu plus éloigné sur le jeu de données RS_Foie_dose3, ces deux méthodes restent les plus similaires. En revanche, la méthode de SVM-RFE se démarque toujours des deux autres : elle ne sélectionne pas les séquences trouvées par les autres méthodes et elle en sélectionne parfois certaines qui lui sont spécifiques (RS_Foie_dose2, RS_Foie_dose3). Pour les autres organes, soit la configuration est identique à RS_Foie_dose2, soit le T-test présente lui aussi des séquences qui lui sont spécifiques.

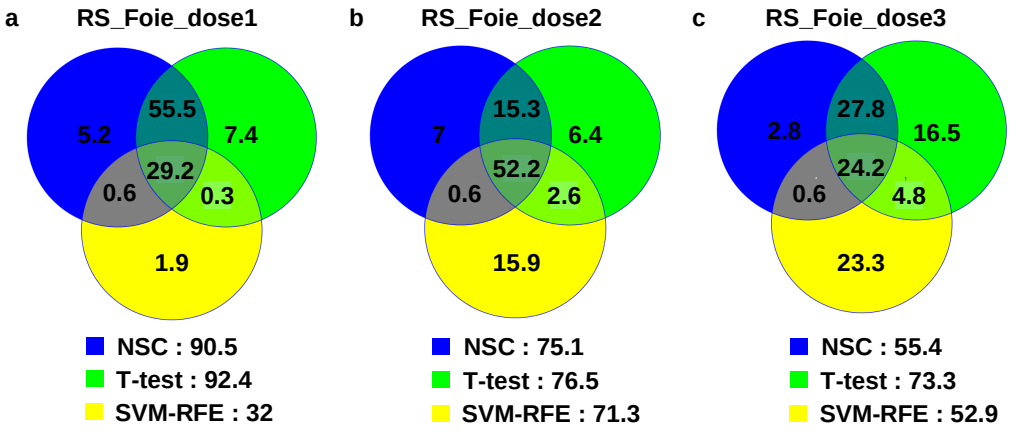

FIG. 4.9 : Comparaison des listes de séquences obtenues avec les trois méthodes de sélection de variables pour toutes les doses dans le foie
Le pourcentage de chaque groupe de séquences a été calculé par rapport à l'ensemble des séquences trouvées toutes méthodes confondues.

4.3.2.2 Validation de l'absence de séquences non exprimées

Dans le cadre de l'étude des séquences sélectionnées, nous avons également voulu vérifier l'absence de séquences non exprimées parmi les résultats. MetRob recherche des séquences différemment régulées entre deux traitements (rosiglitazone et SCOMP) par rapport à une référence commune (souris db/db). Il n'y a néanmoins aucune contrainte sur l'expression des séquences dans au moins un des groupes d'animaux : groupe témoin ou un des groupes de traitement. En effet, le bruit de fond lié à la technologie des puces à ADN implique que les intensités sur une lame ne sont pas nulles. Des séquences non exprimées pourraient donc avoir été sélectionnées par erreur et ce point doit être vérifié.

Pour cela, une information fournie par le logiciel d'analyse d'image Feature extraction a été utilisée. Ce logiciel réalise, pour chaque spot de la lame, un test de significativité du signal par rapport au bruit de fond (voir Annexe B : valeur IsPosAndSignif). Cette information, qui n'est pas reprise dans les combinaisons de lames effectuées par le logiciel Rosetta Resolver, nécessite de réaliser l'étude au niveau des features (et non pas des reporters : cf 1.4.2.1) et de découpler les lames obtenues par dye-swap. Cela signifie qu'au lieu de disposer de 6 animaux par traitement, on obtient 12 lames par traitement : 6 lames où le traité est marqué en Cy3 et 6 lames où le traité est marqué en Cy5. Un feature est dit non exprimé pour un canal (Cy3 ou Cy5) si le spot correspondant échoue au test de significativité pour au moins 7 lames sur 12 sur ce canal. Ce feature est définit comme NE (non exprimé) s'il est non exprimé pour l'intensité des deux canaux dans chaque traitement. Cela correspond à dire qu'un feature est NE s'il est non exprimé à la fois pour les animaux témoins et pour les deux groupes de traitements.

	Lame complète			**Séquences préfiltrées**		
	NbTot	**NbNE**	**Pourcent**	**NbTot**	**NbNE**	**Pourcent**
RS_Soleus_dose1	43379	6845	15.78	8310	1	0.01
RS_Soleus_dose2	43379	7326	16.89	8270	1	0.01
RS_Soleus_dose3	43379	6922	15.96	9208	0	0
RS_Foie_dose1	43379	8857	20.42	9303	1	0.01
RS_Foie_dose2	43379	8609	19.85	9172	1	0.01
RS_Foie_dose3	43379	8671	19.99	11859	0	0
RS_TAI_dose1	43379	5542	12.78	11335	3	0.03
RS_TAI_dose2	43379	5657	13.04	14417	3	0.02
RS_TAI_dose3	43379	4887	11.27	19233	5	0.03

TAB. 4.11 : Pourcentages de features NE (non exprimés) pour une lame complète et pour les séquences pré-filtrées

Le pourcentage est calculé pour tous les jeux de données. NbTot : nombre de features sur la lame. NbNE : Nombre de features non exprimés. Pourcent : Pourcentage de features non exprimés sur la lame.

Les pourcentages de features NE ont été considérés pour une lame complète, pour les séquences pré-filtrées et pour les séquences sélectionnées par MetRob. Sur la lame entière, on observe que les résultats sont assez homogènes par tissu, sans différence sensible en fonction de la dose (Tableau 4.11). Il faut bien sûr noter que l'on considère les features qui sont non exprimés à la fois pour le témoin et pour les deux traitements, mais l'on observe tout de même une différence notable entre les tissus. Cette différence a été vérifiée de la manière suivante. Les pourcentages de séquences NE ont été calculés lame par lame en ne considérant que l'intensité correspondant au groupe témoin db/db. Un T-test confirme la significativité de la différence entre les tissus (Tableau 4.12). Il est effectivement logique que le pourcentage de séquences non exprimées varie en fonction du tissu.

	Moyenne	**p-value**		
		Soleus	**Foie**	**TAI**
Soleus	20.25			
Foie	24	5.65×10^{-11}		
TAI	15.31	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	

TAB. 4.12 : Différences de pourcentages de features NE entre les tissus
Les pourcentages ont été calculés pour chaque lame uniquement par rapport aux intensités de l'échantillon db/db. La moyenne par tissu est donnée dans ce tableau. La p-value correspond à un t-test effectué entre toutes les lames de deux tissus.
Soleus : muscle squelettique, TAI : tissu adipeux inguinal.

En ce qui concerne les séquences pré-filtrées, on observe très peu de features NE (moins de 0.03% : Tableau 4.11). Le filtre de p-value utilisé, même s'il prend en compte la différence d'expression entre témoin et traité et non l'expression dans l'absolu, semble donc suffisant pour éliminer les séquences non exprimées. Enfin, avec MetRob, quelle que soit la méthode de sélection de variables utilisée, aucun feature NE n'est sélectionné. L'absence de sélection de séquences non exprimées est donc bien validée.

4.3.3 Pouvoir discriminant des séquences sélectionnées

L'utilisation de MetRob a permis de réduire considérablement le nombre de séquences à étudier d'un point de vue biologique (Tableau 4.13). Contrairement au cas des données simulées, les séquences réellement discriminantes ne sont pas connues. La qualité des résultats a donc uniquement pu être estimée en vérifiant le pouvoir discriminant des séquences sélectionnées. Pour cela deux approches ont été réalisées : calcul d'erreurs de classification par Leave-One-Out-Cross-Validation (LOOCV) et application d'une méthode de Partial Least Square Discriminant Analysis (PLS-DA) pour mesurer l'amélioration du pouvoir discriminant par la sélection des séquences.

	T-test	NSC	SVM-RFE
RS_Soleus_dose1	623	632	420
RS_Soleus_dose2	563	432	214
RS_Soleus_dose3	438	438	420
RS_Foie_dose1	288	282	100
RS_Foie_dose2	120	118	112
RS_Foie_dose3	516	390	372
RS_TAI_dose1	7	6	122
RS_TAI_dose2	349	126	395
RS_TAI_dose3	671	599	437

TAB. 4.13 : Nombre de séquences sélectionnées par MetRob pour les trois méthodes de sélection de variables et pour chaque jeu de données

4.3.3.1 Erreurs par LOOCV

De même que pour les données simulées, les séquences sélectionnées par MetRob (couplée au T-test, à NSC ou à SVM-RFE) ont été utilisées pour classer les données avec un SVM linéaire. L'erreur par Leave-One-Out-Cross-Validation (LOOCV) a été calculée et comparée à celle obtenue en utilisant les séquences SSR (Tableau 4.14).

	Séquences SSR	Séquences MetRob		
		T-test	NSC	SVM-RFE
RS_Soleus_dose1	0	0	0	0
RS_Soleus_dose2	0	0	0	0
RS_Soleus_dose3	0	0	0	0
RS_Foie_dose1	0	0	0	0
RS_Foie_dose2	0	0	0	0
RS_Foie_dose3	0	0	0	0
RS_TAI_dose1	0.5	0	0	0
RS_TAI_dose2	0.08	0	0	0
RS_TAI_dose3	0.17	0	0	0

TAB. 4.14 : Erreurs par Leave-One-Out-Cross-Validation pour les séquences SSR et pour les séquences sélectionnées par MetRob

Les erreurs par LOOCV obtenues avec les séquences de MetRob sont nulles quels que soient le jeu de données ou la méthode de sélection de variable utilisée. L'erreur de discrimination est également nulle pour les séquences SSR dans le muscle et le foie. En revanche, les jeux de données correspondant au TAI semblent être plus difficiles à classifier, surtout pour la dose la plus faible où l'erreur par LOOCV est de 0.5. La qualité de discrimination des séquences obtenue par MetRob est validée et on retrouve la remarque énoncée pour les données simulées : dans la plupart des cas, le choix du nombre de séquences à conserver ne peut être effectué en se basant sur une erreur de classification nulle.

4.3.3.2 Etude PLS-DA

Le calcul des erreurs par LOOCV ne permet pas de réellement mesurer l'amélioration de la classification par la sélection des séquences. Un autre approche a donc été mise en place en utilisant une méthode d'analyse discriminante, la PLS-DA, à travers le logiciel SIMCA-P+ v12.0 [73]. SIMCA-P+ est un logiciel d'analyse de données multivariées qui permet de réaliser des analyses en composantes principales et des analyses de classification ou de régression. La méthode de PLS-DA est une méthode de classification qui adapte la technique d'analyse en composantes principales (cf 2.2.2.1) de manière à concilier deux objectifs : décrire au mieux l'ensemble des variables explicatives et prédire la classe des observations. Le modèle calculé correspond donc aux composantes remplissant au mieux ces objectifs et au choix du nombre de ces composantes à conserver.

Une analyse en PLS-DA fournit différents scores qui ont été utilisés pour évaluer la qualité du modèle construit. X est la matrice des variables et Y est la classe des observations :

- R2Xcum : variation cumulée de X représentée dans le modèle
- R2Ycum : variation cumulée de Y représentée dans le modèle
- Q2cum : variation cumulée de Y prédite par le modèle (calculée par validation croisée)
- P-value : significativité du modèle

Ces scores sont présentés dans le Tableau 4.15 pour la méthode NSC en considérant différents ensembles de séquences pour créer le modèle : toutes les séquences d'une puce à ADN, les séquences SSR et les séquences sélectionnées par MetRob. Pour le muscle, le foie et les doses 84µmol/kg et 280µmol/kg du tissu adipeux inguinal (TAI), on observe de bons R2Ycum et Q2cum et donc une bonne classification quel que soit l'ensemble de séquences considéré. Les résultats de Q2cum sont néanmoins un peu moins bons avec toutes les séquences d'une lame dans le foie. D'autre part, excepté pour les séquences sélectionnées avec MetRob, les R2Xcum sont plutôt faibles. Ceci signifie qu'il reste beaucoup de variabilité dans les données qui n'est pas liée à la discrimination, et donc des séquences non discriminantes. Enfin, les résultats les plus significatifs sont globalement obtenus avec les séquences MetRob. On peut remarquer le cas de la dose 28µmol/kg dans le TAI : le logiciel n'a pu construire un modèle que pour les séquences MetRob. Cela confirme que ce jeu de données est le plus difficile à

classifier, ce qui est cohérent avec des résultats observés précédemment : non significativité des séquences de ce jeu de données pour une utilisation du T-test avec q-value (cf 3.3.1) et faible nombre de séquences sélectionnées avec le T-test et NSC (Tableau 4.13). Les résultats de PLS-DA valident donc l'apport de la sélection de variables en termes de pouvoir discriminant avec des modèles plus significatifs exploitant au mieux les variations de X.

		28µmol/kg			84µmol/kg			280µmol/kg		
		All	**SSR**	**MetRob**	**All**	**SSR**	**MetRob**	**All**	**SSR**	**MetRob**
Soleus	**NSeq**	41174	7401	632	41174	7388	432	41174	8209	438
	R2Xcum	0.352	0.696	0.874	0.33	0.458	0.895	0.279	0.451	0.93
	R2Ycum	0.997	0.996	0.994	0.999	0.998	0.987	0.992	0.99	0.991
	Q2cum	0.904	0.943	0.97	0.9	0.947	0.984	0.892	0.945	0.988
	p-value	0.02	0.03	2×10^{-5}	4×10^{-3}	2×10^{-4}	8×10^{-9}	4×10^{-5}	2×10^{-6}	2×10^{-9}
Foie	**NSeq**	41174	8430	282	41174	8227	118	41174	10770	390
	R2Xcum	0.262	0.474	0.831	0.255	0.329	0.787	0.334	0.632	0.922
	R2Ycum	0.994	0.999	0.994	0.998	0.991	0.973	0.991	0.993	0.993
	Q2cum	0.677	0.933	0.967	0.7	0.815	0.971	0.827	0.94	0.967
	p-value	0.17	0.05	3×10^{-5}	0.27	0.04	1×10^{-7}	0.02	6×10^{-3}	4×10^{-5}
TAI	**NSeq**	41174	10255	6	41174	13184	126	41174	17622	599
	R2Xcum	NA	NA	0.854	0.476	0.577	0.914	0.49	0.488	0.745
	R2Ycum	NA	NA	0.953	0.995	0.954	0.996	0.999	0.989	0.992
	Q2cum	NA	NA	0.941	0.832	0.819	0.964	0.912	0.846	0.966
	p-value	NA	NA	3×10^{-6}	0.37	0.05	7×10^{-4}	0.51	0.17	4×10^{-5}

TAB. 4.15 : Qualité des modèles obtenus par PLS-DA en fonction de l'ensemble de séquences considéré dans le cas de la méthode NSC

Nseq : Nombre de séquences, Soleus : muscle squelettique, TAI : Tissu adipeux inguinal, All : Toutes les séquences présentes sur une puce à ADN, SSR : séquences statistiquement significativement régulées, MetRob : séquences obtenues avec MetRob couplée à NSC.

4.3.4 Impact de l'ajout d'animaux dans le foie

La dernière étape a consisté à étudier l'impact de l'ajout d'animaux sur les résultats. Comme cela avait été présenté au Chapitre 1 (cf 1.6.1), on dispose de données de puces à ADN pour 18 animaux par dose au lieu de seulement 12 dans le foie. Les 12 animaux de l'étude générique ont essentiellement été sélectionnés à partir de la qualité des ARNs dans le muscle squelettique. Cette sélection n'a donc pas eu d'impact sur la qualité des puces à ADN réalisées pour les foies des animaux supplémentaires. En revanche, pour la dose 28µmol/kg,

les souris supplémentaires traitées par rosiglitazone sont les souris ayant les glycémies les plus faibles après traitement. Ces quelques éléments ayant été précisés, les séquences sélectionnées avec n = 12 ou n = 18 animaux ont donc été comparées.

Pour rappel, une séquence est dite statistiquement significativement régulée (SSR) si son expression pour au moins la moitié des animaux d'un des deux traitements est statistiquement différente de l'expression du groupe témoin. Pour n = 18, il y a 9 souris par traitement et ces séquences peuvent donc être définies de deux manières différentes : la régulation peut être exigée pour 4 ou pour 5 animaux. Au vu du nombre de séquences obtenues (Tableau 4.16), le choix s'est porté sur un seuil de 4, c'est-à-dire un seuil correspondant à la partie entière du nombre d'animaux d'une classe divisé par deux. En termes de pourcentage de séquences communes, ces séquences SSR sont assez proches de celles obtenues pour n=12 : 91.74% pour 28 µmol/kg, 90.67% pour 84 µmol/kg et 93.01% pour 280 µmol/kg.

		28µmol/kg	**84µmol/kg**	**280µmol/kg**
n = 6		8430	8227	10770
n = 9	**Seuil : 4 animaux**	8401	8403	10932
	Seuil : 5 animaux	6852	6817	8994

TAB. 4.16 : Nombre de séquences SSR en fonction du nombre d'observation par classe et du seuil de régulation choisi
Le seuil correspond au nombre d'animaux dans une classe pour lequel une séquence doit être régulée afin d'être considérée comme SSR.

Dans un premier temps, le nombre de séquences obtenues en fonction du nombre d'animaux a été considéré. Le Tableau 4.17 présente le nombre de séquences choisi par maximisation de la fonction Diff (cf 3.4.1), la robustesse qui lui est associée et le nombre final de séquences obtenues avec MetRob. Avec n = 18, moins de séquences sont sélectionnées quelle que soit la méthode, parfois de manière assez radicale. La méthode SVM-RFE présente la seule exception pour la dose 84µmol/kg en sélectionnant plus de séquences avec 18 animaux. En parallèle, cette diminution est associée à une augmentation de la robustesse sur les 2 premières doses, et à une stagnation sur la troisième. Les résultats obtenus avec n = 18 sont donc globalement plus robustes avec moins de séquences. Il faut alors vérifier si ces séquences étaient déjà sélectionnées avec n = 12. Pour cela les listes obtenues avec MetRob pour 12 et 18 animaux ont été comparées (Tableau 4.18).

		Choix d'un nombre de séquences		Robustesse		Séquences finales MetRob	
		n = 12	**n = 18**	**n = 12**	**n = 18**	**n = 12**	**n = 18**
28 µmol/kg	**T-test**	610	410	72.01	76.57	288	224
	NSC	580	360	72.23	77.05	282	203
	SVM-RFE	290	220	63.48	66.48	100	89
84 µmol/kg	**T-test**	260	80	71.09	77.4	120	46
	NSC	250	70	71.97	77.4	118	40
	SVM-RFE	270	600	67.44	70.9	112	264
280 µmol/kg	**T-test**	910	100	76.19	76.28	516	59
	NSC	720	100	76.23	78.51	390	62
	SVM-RFE	890	520	70.4	69.54	372	231

TAB. 4.17 : Comparaison des longueurs de listes de séquences et des robustesses pour n = 12 ou n = 18

Pour les méthodes T-test et NSC aux doses 84µmol/kg et 280µmol/kg, les séquences sélectionnées avec n = 18 sont sensiblement incluses dans celles sélectionnées avec n = 12. A la dose 28µmol/kg, le pourcentage de séquences communes diminue aux alentours de 80%. Les résultats sont donc un peu plus éloignés et cette différence peut être expliquée par la remarque sur la glycémie faite en début de partie. Le cas de la SVM-RFE est à part. Ses résultats sont similaires aux autres méthodes à la dose 28µmol/kg. Cependant, comme cela a été vu précédemment, à la dose 84µmol/kg, la méthode sélectionne plus de séquences avec n = 18. A la dose 280µmol/kg, seules 60% des séquences sont communes entre n = 12 et n = 18. Sur données réelles, SVM-RFE est donc la méthode la moins stable par rapport au nombre d'animaux. Pour les méthodes T-test et NSC, l'augmentation de n réduit uniquement le nombre de séquences sans en sélectionner de nouvelles, peut-être en supprimant des séquences non discriminantes.

	T-test	**NSC**	**SVM-RFE**
28µmol/kg	78.13	81.77	73.03
84µmol/kg	93.48	92.5	75.89
280µmol/kg	96.61	100	57.14

TAB. 4.18 : Pourcentage de séquences communes entre les listes obtenues pour n = 12 ou n = 18
Le pourcentage de séquences communes est calculé par rapport à plus petite longueur des deux listes, c'est-à-dire par rapport aux 18 animaux par classe, sauf à la dose 84µmol/kg pour SVM-RFE.

4.4 Conclusion

Les performances des trois méthodes de sélection de variables ont été comparées et la méthodologie MetRob a été évaluée à partir de jeux de données simulées et de données réelles. En termes de robustesse, la méthode de SVM-RFE est moins robuste pour les deux types de données. Sur données réelles, la méthode NSC est plus robuste et donc plus intéressante que le T-test. D'autre part, les séquences sélectionnées avec le T-test ou la méthode de NSC sont assez proches et la méthode de SVM-RFE se démarque des deux autres pour les données réelles et simulées.

L'intérêt de la méthodologie MetRob réside principalement dans l'obtention d'une bonne valeur prédictive positive (VPP) : toutes les séquences réellement discriminantes ne sont pas sélectionnées, mais les séquences sélectionnées sont fiables. Ces résultats ont été obtenus sur données simulées. En outre, sur données réelles, les séquences sélectionnées par MetRob possèdent un pouvoir discriminant fortement amélioré par rapport aux séquences statistiquement significativement régulées.

En termes d'impact du nombre d'observations, considérer 18 animaux consiste, pour les données PPAR, à sélectionner des séquences déjà choisies pour 12 animaux, mais en plus faible quantité. La méthode de SVM-RFE est de plus la moins stable par rapport à ces variations de nombre d'observations. Les résultats sur données simulées sont plus difficiles à interpréter car ils diffèrent en fonction du jeu de données considéré. On peut néanmoins remarquer que la situation de Sim2 se rapproche des résultats sur données réelles : diminution de la sensibilité avec une conservation de la VPP et une diminution du nombre de séquences pour NSC et T-test et augmentation de la sensibilité et du nombre de séquences pour SVM-RFE. Ce jeu de données correspond à un nombre moyen de séquences discriminantes où les distributions des séquences discriminantes et non discriminantes se recoupent, ce qui semble a priori être proche d'une situation réelle.

Les résultats présentés dans ce chapitre montrent une supériorité de la méthode NSC en termes de robustesse, principalement sur données réelles. Les séquences sélectionnées par le T-test et la NSC étant, dans beaucoup de cas, assez similaires, la méthode de NSC est donc à privilégier pour sa robustesse. D'autre part, la SVM-RFE obtient parfois de meilleures sensibilités, mais elle n'est pas stable sur données réelles par rapport au nombre d'observations. C'est donc sur la méthode NSC que s'est essentiellement focalisée l'étude biologique des résultats présentée dans le chapitre suivant.

Chapitre 5 :

Résultats : Analyse biologique des listes de séquences sélectionnées

L'efficacité des méthodes de sélection de variables et de la méthodologie MetRob a été évaluée du point de vue de la robustesse et de la pertinence des séquences sélectionnées. La méthode de Nearest Shrunken Centroids (NSC) a été choisie pour sa robustesse et sa stabilité afin d'étudier les séquences sélectionnées d'un point de vue biologique. Par la suite, on appellera NSCRob la méthodologie MetRob couplée à la méthode de sélection de variables NSC. Dans le cadre de cette étude, les séquences sélectionnées sont celles qui expliquent au mieux les différences entre deux agonistes PPAR, la rosiglitazone et le SCOMP. Ces deux composés ont été comparés dans le cadre de la recherche de traitements contre le diabète de type 2, maladie affectant le métabolisme. Les séquences obtenues avec NSCRob ont donc principalement été analysées d'un point de vue métabolique.

Les principales voies métaboliques étudiées sont présentées dans une première partie. L'enrichissement des listes de séquences sélectionnées en séquences appartenant aux voies métaboliques a ensuite été considéré. Nous nous sommes pour cela demandé si ces séquences étaient préférentiellement sélectionnées par NSCRob. Enfin, nous nous sommes intéressés au lien entre observations biologiques (paramètres mesurés : glycémie, acides gras, triglycérides) et observations transcriptomiques.

5.1 Détail des principales voies métaboliques

Le métabolisme est l'ensemble des réactions chimiques se produisant dans les cellules de l'organisme. Il permet de dégrader les nutriments et d'en extraire de l'énergie (catabolisme), mais aussi de synthétiser les constituants nécessaires à la structure et au fonctionnement des cellules (anabolisme). Les principales voies du métabolisme des sucres et des graisses qui sont présentées ici peuvent être retrouvées dans Berg et al. [4].

5.1.1 Glycolyse et néoglucogenèse

Le glucose est la principale source d'énergie de la plupart des tissus (muscle, foie, cerveau). Pour pouvoir être utilisé, le glucose est tout d'abord importé dans la cellule via des transporteurs membranaires spécifiques, puis il est dégradé en pyruvate par la voie appelée glycolyse (Figure 5.1).

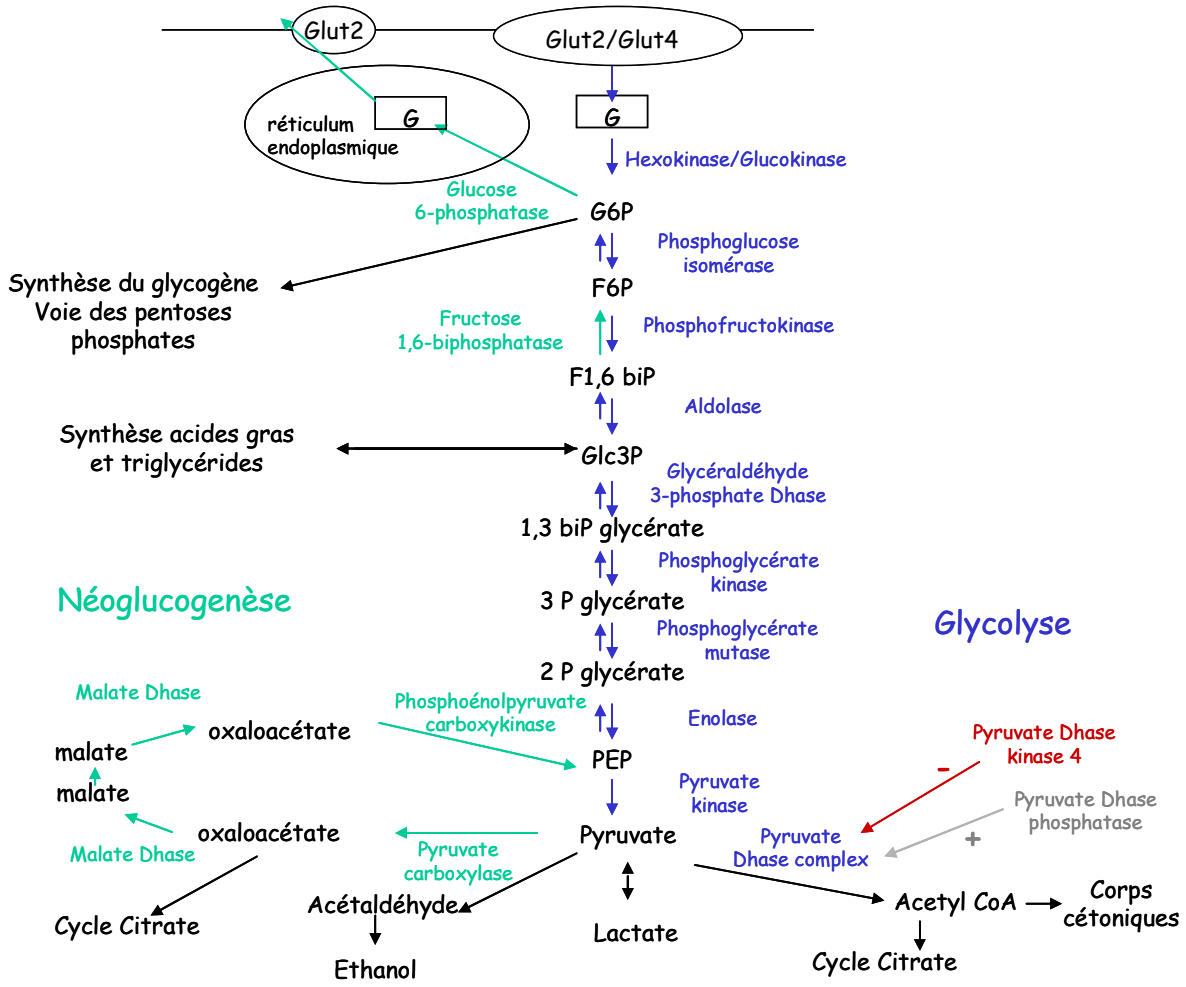

FIG. 5.1 : Voies de la glycolyse et de la néoglucogenèse
Les enzymes en bleu et vert catalysent respectivement des réactions de la glycolyse et de la néoglucogenèse. G : glucose, F : fructose, P : phosphate, biP : biphosphate, Glc : glycéraldéhyde, PEP : phosphoenolpyruvate, CoA : coenzyme A, Dhase : déshydrogénase.

La glycolyse s'effectue dans le cytosol en 3 phases : conversion du glucose en fructose 1,6-biphosphate, clivage du fructose 1,6-biphosphate en 2 fragments tricarbonés (glycéraldéhyde 3-phosphate), puis récupération d'Adénosine TriPhosphate (ATP) lors de l'oxydation des fragments tricarbonés en pyruvate. Elle est principalement contrôlée par trois enzymes qui catalysent les réactions irréversibles du processus : hexokinase, phosphofructokinase et pyruvate kinase. Le glucose 6-phosphate, intermédiaire de la glycolyse, peut également être utilisé pour synthétiser des réserves de glycogène (polymère ramifié constituant une mise en réserve du glucose rapidement mobilisable) ou pour générer du Nicotinamide Adénine Dinucléotide Phosphate réduit (NADPH), indispensable aux réactions réductrices de biosynthèse (voie des pentoses phosphates).

En temps normal, le glucose est formé à partir des réserves de glycogène, mais lors d'une longue période de jeûne, il peut être synthétisé à partir de précurseurs non glucidiques (lactate, aminoacides, glycérol). Cette voie métabolique est appelée néoglucogenèse (Figure 5.1) et se produit principalement dans le foie et dans une moindre mesure dans les reins, le muscle squelettique et le muscle cardiaque. Du fait des trois réactions irréversibles de la glycolyse, la néoglucogenèse n'en est pas la voie inverse. Ces trois étapes sont en effet contournées. La première étape de transformation du pyruvate en oxaloacétate s'effectue dans les mitochondries. Puis l'oxaloacétate est transporté dans le cytosol (sous forme de malate). La fructose 1,6-biphosphatase permet la transformation du fructose 1,6-biphosphate en fructose 6-phosphate. Dans la plupart des tissus, la gluconéogenèse est stoppée au glucose 6-phosphate qui ne peut pas diffuser au-dehors de la cellule et sert en particulier à former du glycogène. La glucose 6-phosphatase qui permet la dernière étape de formation de glucose est majoritairement exprimée et fonctionnelle dans le foie, et présente à un moindre degré dans les reins (organes permettant de réguler la glycémie). La synthèse du glucose s'effectue dans la lumière du réticulum endoplasmique, puis le glucose est de nouveau transporté dans le cytosol pour pouvoir être exporté de la cellule. Les taux et les activités enzymatiques de chaque voie sont contrôlés de façon à ce que glycolyse et néoglucogenèse ne soient pas pleinement actives en même temps.

5.1.2 Cycle du citrate et phosphorylation oxydative

Le pyruvate synthétisé lors de la glycolyse peut ensuite être métabolisé en lactate, en éthanol ou en acétyl coenzyme A (Acétyl CoA). L'Acétyl CoA entre alors dans le cycle du citrate (Figure 5.2) qui est la voie commune terminale d'oxydation des molécules énergétiques (sucres, graisses, protéines). Cet ensemble de réactions s'effectue dans les mitochondries. Le cycle du citrate est un processus aérobie qui permet une oxydation complète des métabolites en dioxyde de carbone et est également source de précurseurs biosynthétiques (acides aminés, acides gras, stérols). Une molécule d'Acétyl CoA produit deux molécules de dioxyde de carbone, une molécule énergétique de Guanosine TriPhosphate (GTP) et 8 électrons. Le gradient de protons induit par ces électrons est ensuite utilisé par la phosphorylation oxydative pour produire de l'énergie via la fabrication de molécules d'ATP. La chaîne

respiratoire est constituée de quatre complexes : trois pompes à protons et un lien physique avec le cycle du citrate.

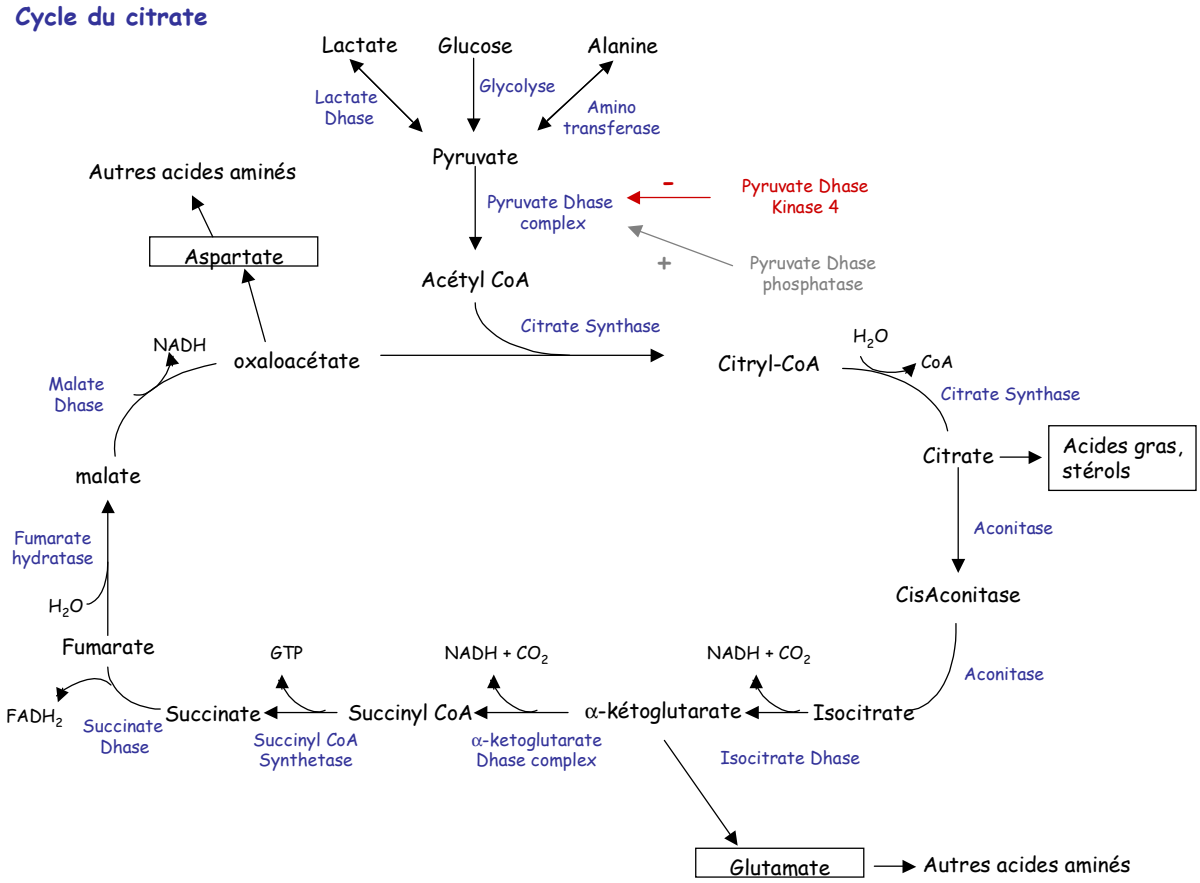

FIG. 5.2 : Cycle du citrate
Dhase : déshydrogénase, CoA : coenzyme A, H_2O : eau, CO_2 : dioxyde de carbone, NADH : nicotinamide adénine dinucléotide réduit, $FADH_2$: flavine adénine dinucléotide réduit, GTP : guanosine triphosphate.

5.1.3 Métabolisme des triglycérides

Le glucose n'est pas la seule source d'énergie de l'organisme. Les lipides, stockés sous forme de triglycérides, constituent une réserve d'énergie extrêmement concentrée. Les triglycérides alimentaires nécessitent une dégradation en acides gras et monoglycérides qui leur permet de passer la barrière intestinale. Ils sont ensuite de nouveau synthétisés dans les cellules de la muqueuse intestinale, puis transportés dans le sang sous forme de chylomicrons, complexes

constitués essentiellement de triglycérides et d'apolipoprotéines. Ces étapes de dégradation et synthèse sont de nouveau nécessaires pour leur import dans les cellules.

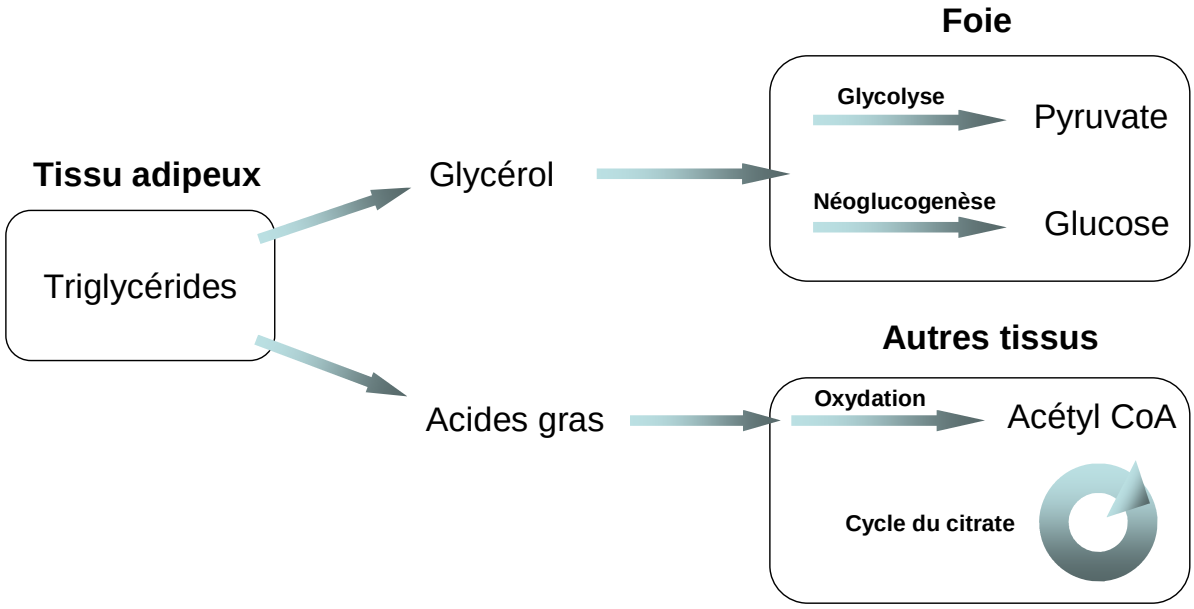

FIG. 5.3 : Utilisation des triglycérides par l'organisme (d'après [4])
CoA : coenzyme A.

Les triglycérides s'accumulent surtout dans les cellules adipeuses qui sont spécialisées dans leur synthèse et leur stockage. Ils peuvent également être stockés par le muscle qui les oxydera pour ses propres besoins énergétiques. Le tissu adipeux permet de diffuser les triglycérides pour fournir de l'énergie aux autres organes. Pour cela, les triglycérides sont transformés en acides gras et glycérol, puis exportés dans le sang vers les autres organes. Le glycérol peut être converti en pyruvate ou en glucose dans le foie et les acides gras peuvent être oxydés dans tous les tissus avant d'entrer dans le cycle du citrate (Figure 5.3).

5.1.4 Métabolisme des acides gras

Les acides gras sont issus de la dégradation des triglycérides et ils peuvent être oxydés pour fournir de l'énergie. Pour cela, ils sont transportés à l'intérieur de la cellule via des transporteurs membranaires, puis ils sont oxydés en Acétyl CoA (Figure 5.4). Il existe trois voies d'oxydation des acides gras : la β-oxydation mitochondriale (la plus courante), la β-oxydation péroxisomale (raccourcissement des acides gras avant la β-oxydation mitochondriale) et l'ω-oxydation (oxydation cytosolique intervenant en cas de déficience de la β-oxydation). L'Acétyl CoA créé entre ensuite dans le cycle du citrate pour fournir de

l'énergie aux cellules. En revanche, s'il y a trop de dégradation des lipides par rapport aux glucides, il est transformé en corps cétoniques qui sont majoritairement produits dans le foie. Ces corps cétoniques peuvent être sources d'énergie pour le muscle cardiaque, le cortex rénal et le cerveau en cas de jeûne. L'Acétyl CoA est également un précurseur du cholestérol, qui peut être synthétisé dans le foie.

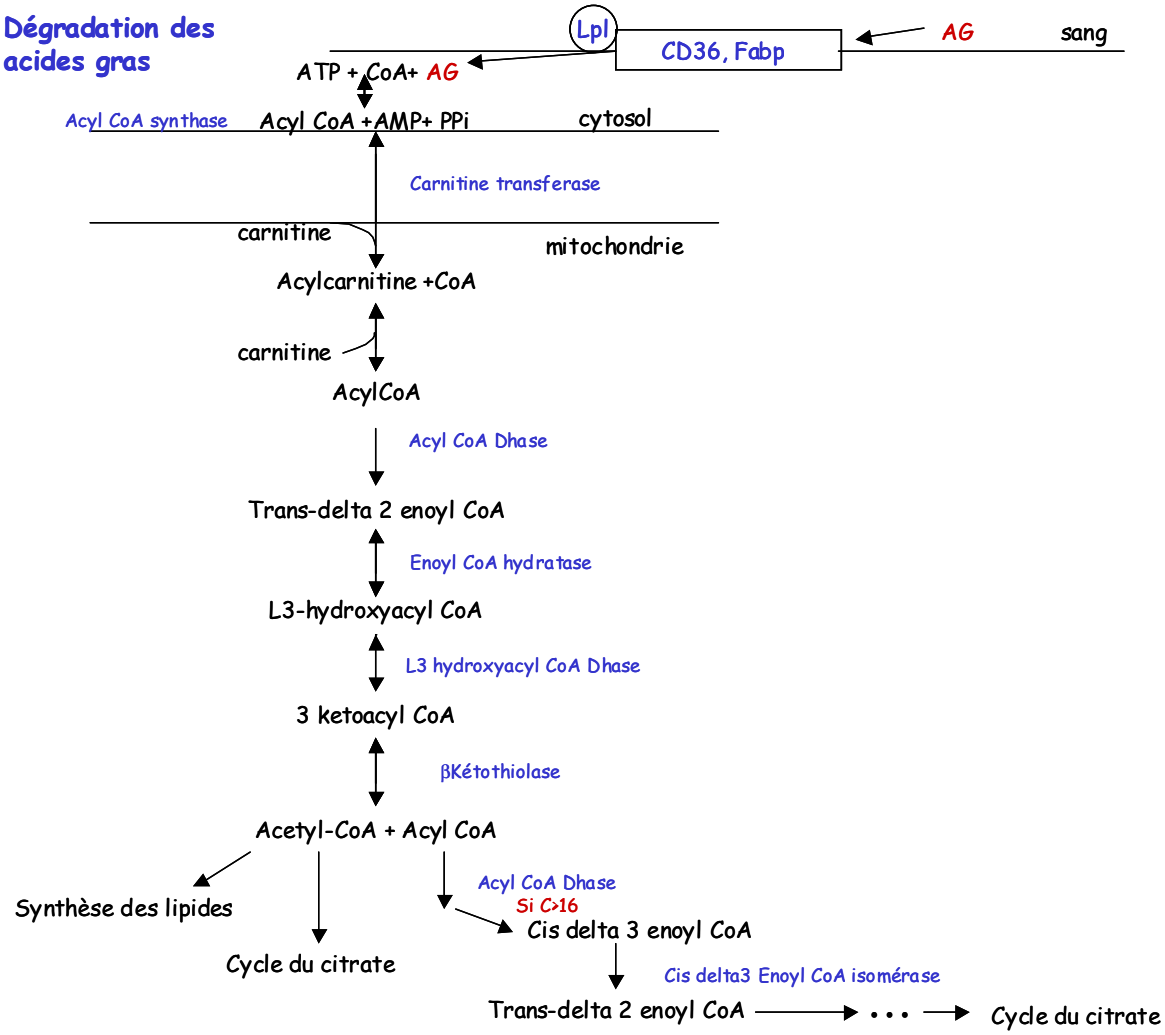

FIG. 5.4 : Dégradation des acides gras : β-oxydation
ATP : adénosine triphosphate, CoA : coenzyme A, AG : acides gras, AMP : adénosine monophosphate, Ppi : pyrophosphate, Dhase : déshydrogénase, Lpl : lipoprotéine lipase, CD36 : antigène CD36, Fabp : protéine de liaison des acides gras.

La synthèse des acides gras s'effectue par les mêmes réactions fondamentales, mais les mécanismes en sont différents : réactions dans le cytosol, liaison des intermédiaires de synthèse à une protéine de transport d'acyles au lieu du coenzyme A, synthèse par un

complexe enzymatique multifonctionnel, ... Parmi les enzymes clés de la synthèse des acides gras, on peut citer le complexe acide gras synthase, l'ATP-citrate lyase et l'acétyl-CoA-carboxylase.

5.1.5 Intégration des voies métaboliques

Les voies métaboliques présentées ci-dessus sont intégrées au sein d'une cellule (Figure 5.5). Elles sont toutes représentées mais leur importance diffère en fonction du type de cellules : musculaires, adipeuses, hépatiques, ... La glycolyse se produit plutôt dans le muscle et le foie, et la néoglucogenèse principalement dans le foie. Les triglycérides sont essentiellement synthétisés dans le tissu adipeux pour y être stockés. Le foie peut également en synthétiser, mais ils sont alors exportés dans le sang via les lipoprotéines de basse densité. La dégradation des triglycérides s'effectue surtout dans le tissu adipeux et les autres types cellulaires traitent essentiellement les produits de cette dégradation : glycérol et acides gras.

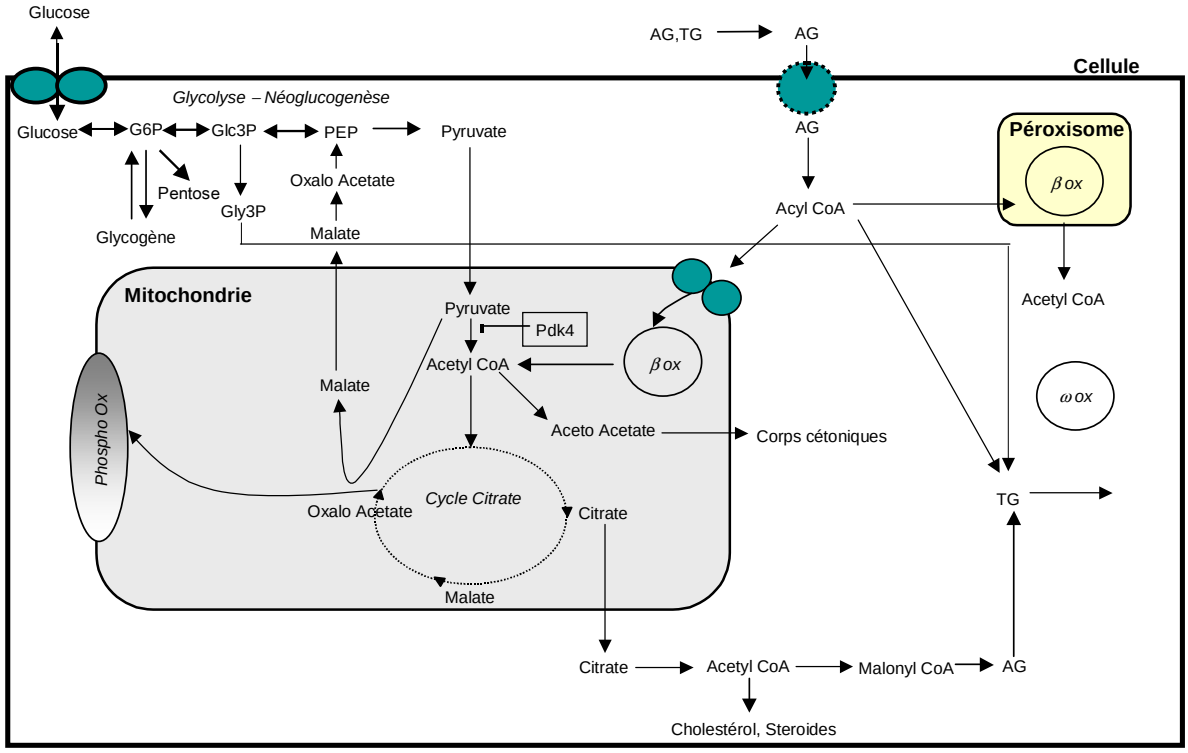

FIG. 5.5 : Schéma de l'intégration des voies métaboliques au sein d'une cellule (d'après [19])
TG : triglycérides, AG: acides gras, G6P : glucose 6-phosphate, Glc3P: glycéraldéhyde 3-phosphate, Gly3P : glycérol 3-phosphate, PEP: phosphoénolpyruvate, CoA : coenzyme A, ox : oxydation, Pdk4 : pyruvate déshydrogénase kinase 4.

5.2 Enrichissements en voies métaboliques des annotations associées aux listes de séquences sélectionnées

Les deux composés comparés dans notre étude sont des agonistes PPAR ciblant le diabète de type 2. PPAR régule principalement des gènes impliqués dans le métabolisme et dans les processus inflammatoires. Nous nous sommes ici concentrés sur la partie métabolique et avons voulu vérifier si les séquences appartenant aux principales voies métaboliques avaient été préférentiellement sélectionnées. Dans cette optique, un enrichissement en séquences des gènes des voies métaboliques a été calculé pour les listes sélectionnées par NSCRob par rapport à une lame de puce à ADN complète.

Dix-huit voies métaboliques d'intérêt totalisant 302 gènes (soit 501 séquences) ont été sélectionnées (Tableau 5.1). Ces voies ne sont pas exhaustives mais contiennent des gènes pour lesquels les fonctions sont consolidées. Le nombre total de séquences et le nombre de séquences dans les voies métaboliques d'intérêt ont été calculés pour la lame entière et pour les listes de séquences sélectionnées par NSCRob. Un test exact de Fisher (voir détail en Annexe D) a ensuite été réalisé afin de déterminer si l'enrichissement obtenu était statistiquement significatif.

Voies métaboliques	**Nombre de gènes**	**Nombre de séquences**
Glycolyse	21	41
Néoglucogenèse	6	8
Métabolisme du glycogène	29	53
Voie des pentoses phosphates	7	10
Métabolisme du pyruvate	28	45
Cycle du citrate	18	27
Phosphorylation oxydative	47	61
Dépenses d'énergie	2	4
Transport des lipides	21	37
Synthèse des acides gras	12	20
Oxydation des acides gras - Transport des acides gras	11	21
Oxydation des acides gras - β-oxydation mitochondriale	19	28
Oxydation des acides gras - β-oxydation peroxisomale	10	13
Oxydation des acides gras - ω-oxydation	42	65
Synthèse du cholestérol	9	20
Synthèse des corps cétoniques	11	19
Synthèse des triglycérides	11	25
Dégradation des triglycérides	6	16

TAB. 5.1 : Voies métaboliques étudiées

5.2.1 Comparaison entre rosiglitazone et SCOMP

Ces enrichissements ont été calculés pour les jeux de données qui permettent la comparaison entre rosiglitazone et SCOMP à trois doses (28 µmol/kg, 84 µmol/kg et 280 µmol/kg) dans trois organes (muscle squelettique, foie et tissu adipeux inguinal) (Tableau 5.2).

		28 µmol/kg	84 µmol/kg	280 µmol/kg
Lame entière	**Nombre de séquences**	41174	41174	41174
	Nombre de séquences des voies métaboliques	501	501	501
	Pourcentage	1.2	1.2	1.2
Soleus	**Nombre de séquences**	632	432	438
	Nombre de séquences des voies métaboliques	8	6	4
	Pourcentage	1.3	1.4	0.9
	P-value	**0.85**	**0.66**	**0.82**
Foie	**Nombre de séquences**	282	118	390
	Nombre de séquences des voies métaboliques	17	17	38
	Pourcentage	6	14.4	9.7
	P-value	**< 0.01**	**< 0.01**	**< 0.01**
TAI	**Nombre de séquences**	6	126	599
	Nombre de séquences des voies métaboliques	0	25	6
	Pourcentage	0	19.8	1
	P-value	**1**	**< 0.01**	**0.85**

TAB. 5.2 : Enrichissements en séquences des voies métaboliques pour les résultats de NSCRob dans la comparaison entre rosiglitazone et SCOMP
Les résultats sont présentés pour tous les organes et toutes les doses de traitement. La p-value représente la significativité de l'enrichissement des listes de séquences sélectionnées par NSCRob par rapport à la lame entière. Les cases sont en vert si la p-value est significative (< 0.01), en rouge sinon. Soleus : muscle squelettique. TAI : tissu adipeux inguinal.

Dans le muscle squelettique, l'enrichissement en séquences des voies métaboliques d'intérêt n'est significatif à aucune dose. Cette absence d'enrichissement pourrait être expliquée par une activité métabolique moindre que dans le foie et le tissu adipeux inguinal (TAI), par une expression faible de PPARγ comparée à celle du TAI (Tableau 5.3) ou par l'activation de mécanismes indépendants du transcriptome.

	Muscle squelettique	**Foie**	**Tissu adipeux inguinal**
PPARα	2134	10149	518
PPARγ	1547	3777	28897

TAB. 5.3 : Niveau d'expression de PPARα et PPARγ dans les organes étudiés
Les valeurs fournies sont les moyennes des intensités correspondant aux souris diabétiques db/db tous jeux de données confondus.

Dans le foie, l'enrichissement est significatif à toutes les doses avec des pourcentages de séquences des voies métaboliques d'intérêt valant respectivement 6, 14.4 et 9.7 pour les doses 28, 84 et 280 µmol/kg (Tableau 5.2). On observe que ce pourcentage décroît à la dose la plus élevée, ce qui est étonnant car une augmentation avec la dose pourrait être attendue. Cependant, la dose 280 µmol/kg est très élevée et les écarts les plus importants entre les deux produits peuvent alors être moins liés au métabolisme et plus à des effets secondaires. Des différences importantes en termes de voies métaboliques sont observées dans le foie, alors que PPARγ y est à peine plus exprimé que dans le muscle squelettique. Cette observation est certainement liée à la composante α du composé SCOMP (voie Annexe A) car PPARα est majoritairement exprimé dans le foie (Tableau 5.3).

Dans le TAI, pour la dose 28 µmol/kg, aucune séquence sélectionnée par NSCRob n'appartient aux voies métaboliques d'intérêt et l'enrichissement n'est pas significatif (Tableau 5.2). Ce résultat n'est pas surprenant car la liste obtenue par NSCRob ne contient que 6 séquences. Il semble y avoir très peu de différences à cette dose dans le TAI entre les deux composés (cf observation faite au chapitre 4, paragraphes 4.3.3.1 et 4.3.3.2). L'enrichissement est significatif à la dose 84 µmol/kg : 19.8% de séquences des voies d'intérêt sont présentes dans la liste obtenue avec NSCRob. C'est le pourcentage le plus important observé ici et il très certainement dû à une expression élevée de PPARγ dans le TAI. En revanche, à la dose 280 µmol/kg, aucun enrichissement significatif n'est observé. Ce phénomène est plus accentué que la diminution observée à la dose la plus importante dans le foie. Outre la possibilité de sélection de séquences liées à des effets secondaires, ce cas est celui qui comporte le plus de variables initiales (17622). Etant donné le faible nombre d'observations considéré jusqu'à présent, les limites des méthodes sont peut-être atteintes pour ce jeu de données.

5.2.2 Impact de l'ajout d'animaux

La comparaison effectuée au chapitre 4 (cf 4.3.4) entre n = 12 et n = 18 observations a permis d'établir que, pour NSCRob, considérer plus d'animaux réduit uniquement le nombre de séquences sans en sélectionner de nouvelles, peut-être en supprimant des séquences non

discriminantes. Cette comparaison a été poursuivie en termes d'enrichissement des séquences sélectionnées en voies métaboliques d'intérêt (Tableau 5.4).

Dans tous les jeux de données, on observe une chute en termes de quantité de séquences du métabolisme en passant de n = 12 à n = 18, mais cette chute est associée à une augmentation du pourcentage d'enrichissement. Cette augmentation est particulièrement flagrante pour la dose 280µmol/kg avec un passage de 9.7% pour n = 12 à 27.4% pour n = 18. On peut alors dans ce cas observer une augmentation de l'enrichissement avec la dose de traitement. Les diminutions d'enrichissement aux hautes doses semblent donc pouvoir être contrebalancées par une augmentation du nombre d'animaux considérés. La dose 280 µmol/kg est celle qui, pour le foie, contient le plus de variables (10770). Il semble donc plus facile de gérer cette quantité de variables avec plus d'observations. Ces résultats appuient l'hypothèse émise au chapitre 4 selon laquelle augmenter le nombre d'observations permettraient d'éliminer des séquences non discriminantes et de conserver les séquences les plus intéressantes.

		28 µmol/kg	**84 µmol/kg**	**280 µmol/kg**
Lame entière	**Nombre de séquences**	41174	41174	41174
	Nombre de séquences des voies métaboliques	501	501	501
	Pourcentage	1.2	1.2	1.2
Foie n = 12	**Nombre de séquences**	282	118	390
	Nombre de séquences des voies métaboliques	17	17	38
	Pourcentage	6	14.4	9.7
	P-value	**< 0.01**	**< 0.01**	**< 0.01**
Foie n = 18	**Nombre de séquences**	206	40	62
	Nombre de séquences des voies métaboliques	16	9	17
	Pourcentage	7.8	22.5	27.4
	P-value	**< 0.01**	**< 0.01**	**< 0.01**

TAB. 5.4 : Enrichissements en séquences des voies métaboliques dans le foie pour les résultats de NSCRob – Comparaison entre n = 12 et n = 18

Les résultats sont présentés pour toutes les doses de traitement. La p-value représente la significativité de l'enrichissement des listes de séquences sélectionnées par NSCRob par rapport à la lame entière. Les cases sont en vert si la p-value est significative (< 0.01), en rouge sinon.

5.2.3 Conclusion

Pour la comparaison entre rosiglitazone et SCOMP, les résultats d'enrichissement observés sont cohérents avec les niveaux d'expression de PPAR dans les tissus et avec les connaissances actuelles du rôle de PPAR dans le contrôle du métabolisme. D'autre part, la comparaison entre n = 12 et n = 18 semble confirmer que l'on conserve les séquences les plus intéressantes quand on augmente le nombre d'observations. Il faut néanmoins rappeler que beaucoup de modifications transcriptomiques ne sont pas directement liées aux gènes cibles de PPAR et peuvent être beaucoup plus difficiles à interpréter que des changements d'ordre métabolique.

5.3 Lien entre observations biologiques et transcriptomiques

Dans l'optique de l'étude des modifications métaboliques, il est également intéressant de mettre en parallèle les observations faites au niveau transcriptomique et les variations des paramètres biologiques observés. Pour rappel, la glycémie a été mesurée à partir de prélèvements sanguins. En parallèle une analyse en lipidomique a été menée sur les souris de l'expérience PPAR. Les quantités de 11 acides gras et de 4 triglycérides ont été mesurées par UPLC-MS (Ultra Performance Liquid Chromatography-tandem Mass Spectrometric) dans le plasma, le foie et le TAI. Les dosages sont des concentrations pour le plasma et ils ont été normalisés par le poids de l'organe pour le foie et le TAI. Au final, trois acides gras (acides palmitique, linoléique et oléique) et deux triglycérides (trilinoléate et trioléate) majoritaires ont été conservés. Les dosages sont présentés en Annexe A pour la dose 84 µmol/kg.

Les modifications biologiques, puis les modifications transcriptomiques ont été étudiées en recherchant les liens qui les unissent pour les souris non diabétiques db+ et pour les souris traitées par rosiglitazone ou SCOMP à la dose 84µmol/kg. Les séquences sélectionnées par NSCRob ont ensuite été considérées. Ces séquences sont censées représenter au mieux les différences entre rosiglitazone et SCOMP. Nous avons donc voulu vérifier si elles permettaient d'expliquer les principales variations biologiques observées.

5.3.1 Modifications des paramètres biologiques

Les paramètres biologiques (glycémie, acides gras, triglycérides) mesurés chez les souris non diabétiques db+ et chez les souris diabétiques db/db traitées à la rosiglitazone ou au SCOMP (dose 84 µmol/kg) ont été comparés à ceux qui ont été observés chez les souris diabétiques db/db (Tableau 5.5). La glycémie est nettement plus basse chez les souris db+ que chez les souris db/db. Cette glycémie plus élevée chez les souris db/db est une des principales caractéristiques du diabète de type 2 [15] et est en partie normalisée par les deux traitements

avec un effet plus important pour la rosiglitazone. Les acides gras et les triglycérides plasmatiques sont présents en moins grande quantité chez les souris non diabétiques. De même que pour la glycémie, rosiglitazone et SCOMP normalisent ces paramètres à la baisse chez les souris db/db. Cette baisse se révèle néanmoins plus importante que nécessaire, surtout avec le traitement par rosiglitazone.

		db+	db/db RGZ	db/db SCOMP
Plasma	Glycémie	↓↓↓	↓↓	↓↓
	AG	↓	↓↓	↓
	TG	↓↓	↓↓↓	↓↓
Foie	AG	↓	↑	↑↑
	TG	↓↓↓	↑	
TAI	AG	↑	↑	↑↑
	TG	↑↑	↑↑	↑↑↑

TAB. 5.5 : Synthèse des modifications biologiques par rapport aux souris diabétiques db/db
db+ : souris non diabétiques, RGZ : rosiglitazone, AG : acides gras, TG : triglycérides, TAI : tissu adipeux inguinal. Les flèches vers le haut correspondent à une augmentation et les flèches vers le bas à une diminution. La quantité de flèches indique l'ampleur du phénomène. Une flèche noire montre la tendance dans les cas où tous les acides gras ou triglycérides n'évoluent pas de la même manière.

Dans le foie, on observe des densités d'acides gras légèrement plus basses chez les souris db+ que chez les souris db/db et des densités de triglycérides très diminuées (Tableau 5.5). Globalement, rosiglitazone et SCOMP augmentent les graisses chez les souris db/db et empirent donc la situation par rapport à l'état non pathologique. Après traitement des souris db/db, une plus forte augmentation de la densité d'acides gras est observée avec le SCOMP, alors que la rosiglitazone augmente plus la densité de triglycérides. Cette augmentation des triglycérides dans le foie (également appelée stéatose) est un effet bien connu de la rosiglitazone chez les rongeurs [74][75].

Au niveau du TAI, les densités d'acides gras et de triglycérides sont globalement augmentées chez les souris db+ par rapport aux souris diabétiques db/db (Tableau 5.5). L'augmentation de la densité de triglycérides y est plus particulièrement marquée, mais la quantité globale de tissu adipeux est diminuée (voir Annexe A). Cette observation suggère un fonctionnement différent des cellules adipeuses et plus particulièrement la présence d'adipocytes plus fonctionnels chez les souris db+. L'effet des deux agonistes PPAR va dans le sens de la

normalisation et ces valeurs de densité sont augmentées pour les souris traitées. Cette augmentation dépasse néanmoins les niveaux des souris db+, surtout pour le SCOMP. De telles observations ne sont pas étonnantes car les thiazolidinediones sont connues pour leur induction de la différentiation adipocytaire et du stockage des triglycérides [76]. La plus grande capacité de stockage des nouveaux adipocytes, plus petits et plus sensibles à l'insuline, permet vraisemblablement d'exercer un meilleur contrôle de l'homéostasie des acides gras plasmatiques.

5.3.2 Modifications transcriptomiques

5.3.2.1 Principe général

Nous avons ensuite considéré les modifications transcriptomiques globales (par voie métabolique : Tableau 5.1) qui ont été observées. Les résultats présentés ci-après comparent le profil transcriptomique des souris diabétiques db/db par rapport aux souris normales non diabétiques db+ et le profil des souris db/db traitées à la dose 84µmol/kg de rosiglitazone par rapport aux souris db/db non traitées dans le foie (Figures 5.6 et 5.7) et dans le TAI (Figures 5.8 et 5.9). Le cas de référence diffère en fonction de la figure car nous avons souhaité mettre en avant les modifications intervenant dans le cas pathologique par rapport à des souris saines : référence db+ pour l'étude des souris db/db et référence db/db pour l'étude des souris db/db traitées à la rosiglitazone. Peu de modifications étant observables dans le muscle squelettique, les observations pour ce tissu ne sont pas développées. D'autre part, les résultats du composé SCOMP ne sont pas présentés pour des raisons de clarté. En effet, même s'il existe des différences sur lesquelles nous reviendront plus tard, les modifications transcriptomiques sont globalement comparables à celles observées pour la rosiglitazone.

La glycolyse et la néoglucogenèse ont été traitées étape par étape avec une attention particulière pour les enzymes limitantes. Dans ce cas, les gènes correspondant aux enzymes de chaque étape ont été considérés pour déterminer la régulation potentielle de la voie. Pour les autres voies métaboliques, nous avons défini des notions d' « up-régulation » et de « down-régulation » potentielles à partir de l'expression des gènes, de manière à pouvoir visualiser les résultats. Ce sont des hypothèses qui nécessiteront d'être vérifiées par d'autres techniques avant toute affirmation. Une voie métabolique a été définie comme potentiellement « up-régulée » si plus de 50% des séquences la composant étaient régulées et si plus de 75% de ces séquences étaient « up-régulées », c'est-à-dire plus exprimés chez le groupe d'intérêt par rapport au groupe de référence. La notion de « down-régulation » a été définie de manière symétrique. Les autres voies ont été considérées comme non affectées. Ces notions ont été représentées sur les figures de cette partie en rouge pour l' « up-régulation » et en vert pour la « down-régulation ».

5.3.2.2 Modifications transcriptomiques dans le foie

Au niveau du glucose, dans le foie, les gènes impliqués dans la néoglucogenèse sont plus exprimés chez les souris db/db que chez les souris db+, de même que les transporteurs de glucose dans le sang (Figure 5.6). On peut en effet citer le transporteur Glut2 (+36%) ainsi que les gènes codants pour la Glucose 6-phosphatase (G6pc : +108%) et la Fructose 1,6-biphosphatase (Fbp1 : +99%), enzymes clés de la néoglucogenèse (cf 5.1.1). Ces observations suggèrent une augmentation de la néoglucogenèse chez les souris db/db, hypothèse confirmée par la littérature [77]. De plus, la glycolyse est possiblement diminuée avec une baisse de sa première étape et un blocage au niveau de la pyruvate déshydrogénase kinase 4 (Pdk4 : +265%), enzyme clé limitante de la glycolyse [78]. Ces observations pourraient contribuer à expliquer l'augmentation de la glycémie observée chez les souris diabétiques.

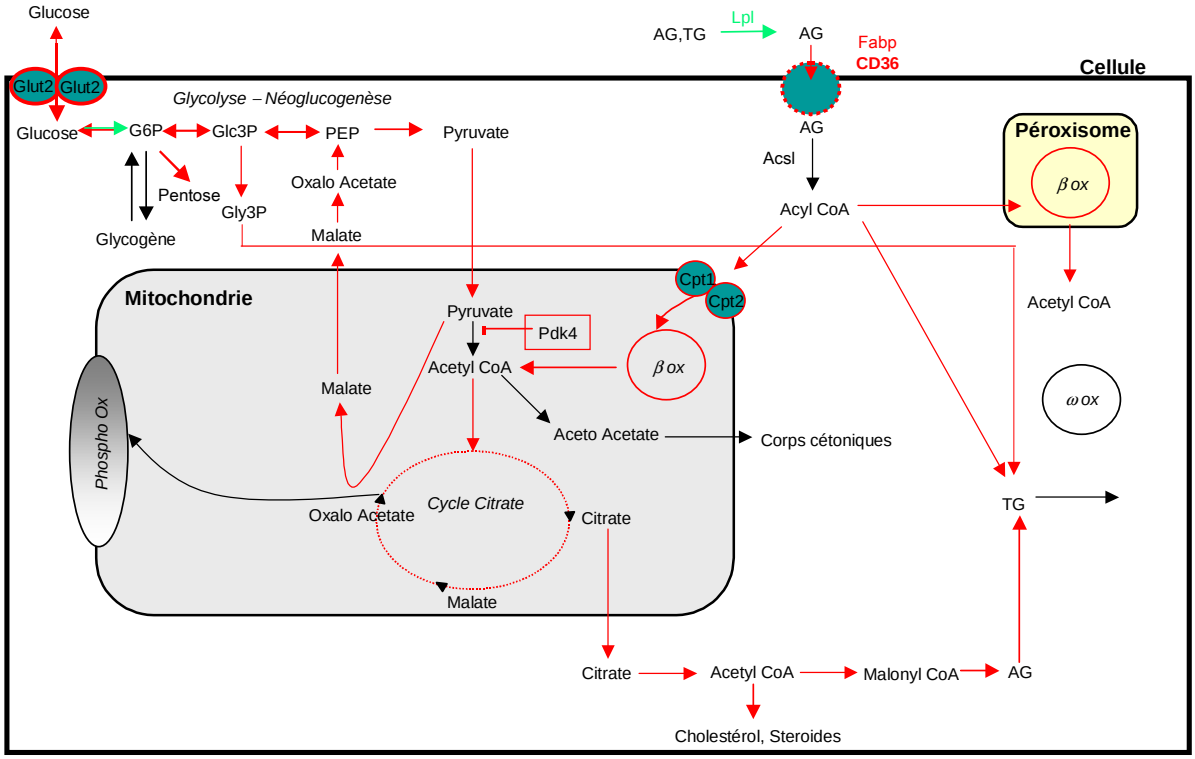

FIG. 5.6 : Modifications transcriptomiques observées dans le foie pour les souris diabétiques db/db par rapport aux souris non diabétiques db+

TG : triglycérides, AG: acides gras, G6P : glucose 6-phosphate, Glc3P: glycéraldéhyde 3-phosphate, Gly3P : glycérol 3-phosphate, PEP: phosphoénolpyruvate, CoA : coenzyme A, ox : oxydation. Glut : transporteur de glucose, Lpl : lipoprotéine lipase, Fabp : protéine de liaison d'acides gras, CD36 : antigène CD36, Acsl : Acyl coenzyme A synthéthase à longue chaîne, Pdk4 : Pyruvate déshydrogénase kinase 4, Cpt : carnitine palmitoyltransférase.

Avec traitement à la rosiglitazone, les gènes impliqués dans la néoglucogenèse déjà cités et les transporteurs de glucose dans le sang sont moins exprimés que chez les souris db/db non traitées. Cependant les gènes impliqués dans la glycolyse ne semblent pas rétablis et Pdk4 (+121%) est encore plus exprimé (Figure 5.7). On peut donc émettre l'hypothèse d'une normalisation de la néoglucogenèse par la rosiglitazone, hypothèse précédemment avancée par Loffler et al. [79]. De nouveau, ces résultats sont cohérents avec les observations biologiques : diminution de la glycémie après traitement.

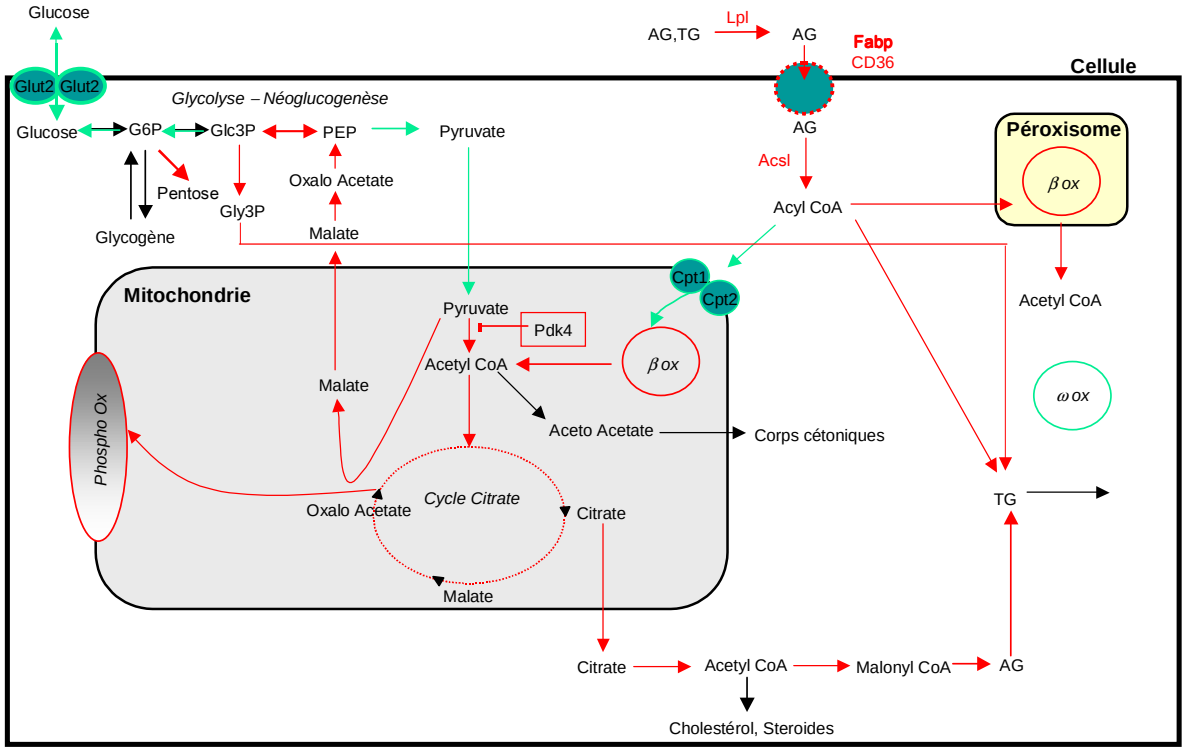

FIG. 5.7 : Modifications transcriptomiques observées dans le foie pour les souris db/db traitées à la rosiglitazone par rapport aux souris db/db non traitées

TG : triglycérides, AG: acides gras, G6P : glucose 6-phosphate, Glc3P: glycéraldéhyde 3-phosphate, Gly3P : glycérol 3-phosphate, PEP: phosphoénolpyruvate, CoA : coenzyme A, ox : oxydation. Glut : transporteur de glucose, Lpl : lipoprotéine lipase, Fabp : protéine de liaison d'acides gras, CD36 : antigène CD36, Acsl : Acyl coenzyme A synthéthase à longue chaîne, Pdk4 : Pyruvate déshydrogénase kinase 4, Cpt : carnitine palmitoyltransférase.

Au niveau des graisses, on observe dans le foie des souris db/db (Figure 5.6) une augmentation des ARNs codant pour les protéines impliquées dans la synthèse des acides gras et des triglycérides telle que le complexe protéique clé de la synthèse des acides gras, l'acide gras synthase (Fasn : +184%) (cf 5.1.4). Il semble également y avoir plus d'import de lipides

dans la cellule, essentiellement des acides gras, avec une sur-expression du transporteur de lipides CD36 (+1082%) et de la protéine de liaison d'acides gras Fabp2 (+64%) chez les souris db/db (cf 5.1.4). De manière potentiellement compensatoire, les gènes impliqués dans les β-oxydations mitochondriale et péroxisomale sont plus exprimés. Néanmoins, cette activation de lipides ne semble pas suffisante car on observe dans le foie et le plasma des souris db/db une augmentation des triglycérides et des acides gras par rapport aux souris db+.

L'administration de rosiglitazone augmente les ARNs codant pour les protéines impliquées dans tous les processus précédents, mais en plus diminue ceux codant pour les protéines impliquées dans l'ω-oxydation (Figure 5.7). Globalement, triglycérides et acides gras sont augmentés par rapport aux souris db/db dans le foie avec le traitement. Il n'y a donc toujours pas assez de compensation par l'oxydation des graisses et/ou il y a plus d'import de lipides (diminution dans le plasma). Cette hypothèse de transfert des graisses entre foie et plasma se retrouve dans la littérature, puisque Watkins et al. ont suggéré que les échanges de lipides entre le foie et le plasma étaient déréglés par la rosiglitazone chez les rongeurs [74].

5.3.2.3 Modifications transcriptomiques dans le tissu adipeux inguinal

Dans le TAI, le métabolisme des sucres est beaucoup moins touché que dans le foie. Nous nous sommes donc plus particulièrement intéressés aux graisses. Comparées aux souris db+, les souris db/db présentent une sous-expression des gènes impliqués dans la synthèse des acides gras et dans les β-oxydations mitochondriale et péroxisomale et une sur-expression des gènes impliqués dans l'import des lipides et l'ω-oxydation (Figure 5.8). Plus particulièrement, l'ATP-citrate lyase (Acly : -77%), l'acétyl-CoA-carboxylase (Acaca : -78%) et Fasn (-85%), protéines impliquées dans la synthèse des acides gras (cf 5.1.4), sont moins exprimés chez les souris db/db. Il en est de même pour l'énoyl-CoA-hydratase (Echs1 : -43%), l'acyl-CoA-déshydrogénase à très longue chaîne (Acadvl : -34%) et l'acétyl-CoA-acyl-transférase (Acaa1b : -24%), enzymes des β-oxydations mitochondriale et péroxisomale (cf 5.1.4). En parallèle, cela correspond des densités de triglycérides et d'acides gras plus faibles dans le TAI, bien que la masse totale de tissu adipeux soit supérieure.

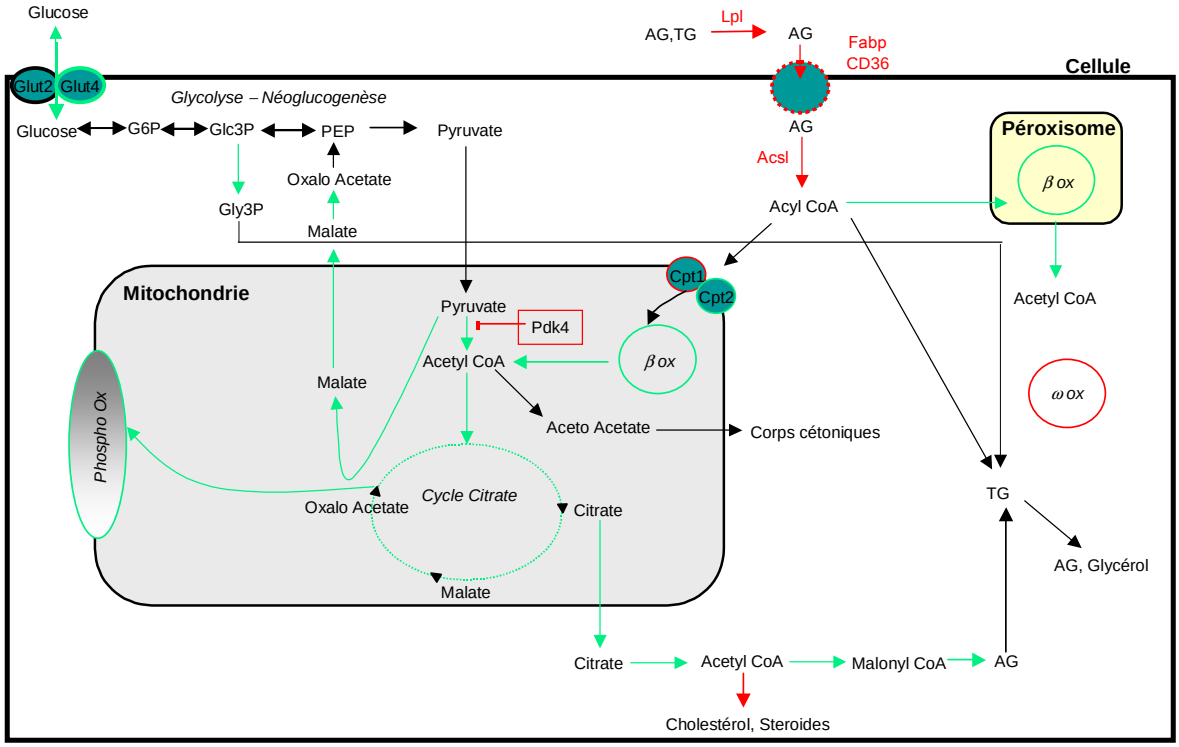

FIG. 5.8 : Modifications transcriptomiques observées dans le tissu adipeux inguinal pour les souris diabétiques db/db par rapport aux souris non diabétiques db+

TG : triglycérides, AG: acides gras, G6P : glucose 6-phosphate, Glc3P: glycéraldéhyde 3-phosphate, Gly3P : glycérol 3-phosphate, PEP: phosphoénolpyruvate, CoA : coenzyme A, ox : oxydation. Glut : transporteur de glucose, Lpl : lipoprotéine lipase, Fabp : protéine de liaison d'acides gras, CD36 : antigène CD36, Acsl : Acyl coenzyme A synthéthase à longue chaîne, Pdk4 : Pyruvate déshydrogénase kinase 4, Cpt : carnitine palmitoyltransférase.

L'administration de rosiglitazone augmente dans le TAI à la fois les ARNs codant pour les protéines impliquées dans la synthèse des triglycérides et dans leur dégradation pour export. Les gènes impliqués dans la synthèse des acides gras sont également plus exprimés après traitement, ainsi que ceux impliqués dans les β-oxydations mitochondriale et péroxisomale (Figure 5.9). En termes de paramètres biologiques, la rosiglitazone tend à la normalisation en augmentant les densités de triglycérides et d'acides gras dans le TAI. Le lien entre transcriptomique et observations biologiques est ici plus difficile à établir. On peut surtout remarquer que la tendance est à la normalisation sur les deux niveaux après traitement par rosiglitazone.

Si l'on se place au niveau de l'activité mitochondriale, on peut néanmoins remarquer que les ARNs codant pour les protéines impliquées dans le cycle du citrate et la phosphorylation oxydative sont diminués chez les souris db/db (Figure 5.8), mais réaugmentés par le traitement (Figure 5.9). Par exemple, la fumarate hydratase (Fh1 : -36%) et la succinate déshydrogénase (Sdha : -40%), ainsi que les membres du complexe III de la phosphorylation

oxydative Uqcrc2 (-39%) et Uqcrfs1 (-36%) sont sous-exprimés chez les souris db/db, mais leur expression est rétablie chez les souris traitées (cf 5.1.2). Cette normalisation par la rosiglitazone de très nombreux gènes impliqués dans le cycle du citrate et la phosphorylation oxydative est bien représentée sur le damier d'expression de la Figure 5.10. Cela suggère une augmentation de l'activité mitochondriale après traitement, observation en accord avec le rétablissement de la genèse mitochondriale précédemment observé chez des souris db/db traitées à la rosiglitazone [80]. Cette normalisation semble donc correspondre à l'apparition, déjà référencée dans la bibliographie pour les chiens et les rongeurs, d'adipocytes plus fonctionnels, plus petits et possédant plus de mitochondries dans le tissu adipeux [81].

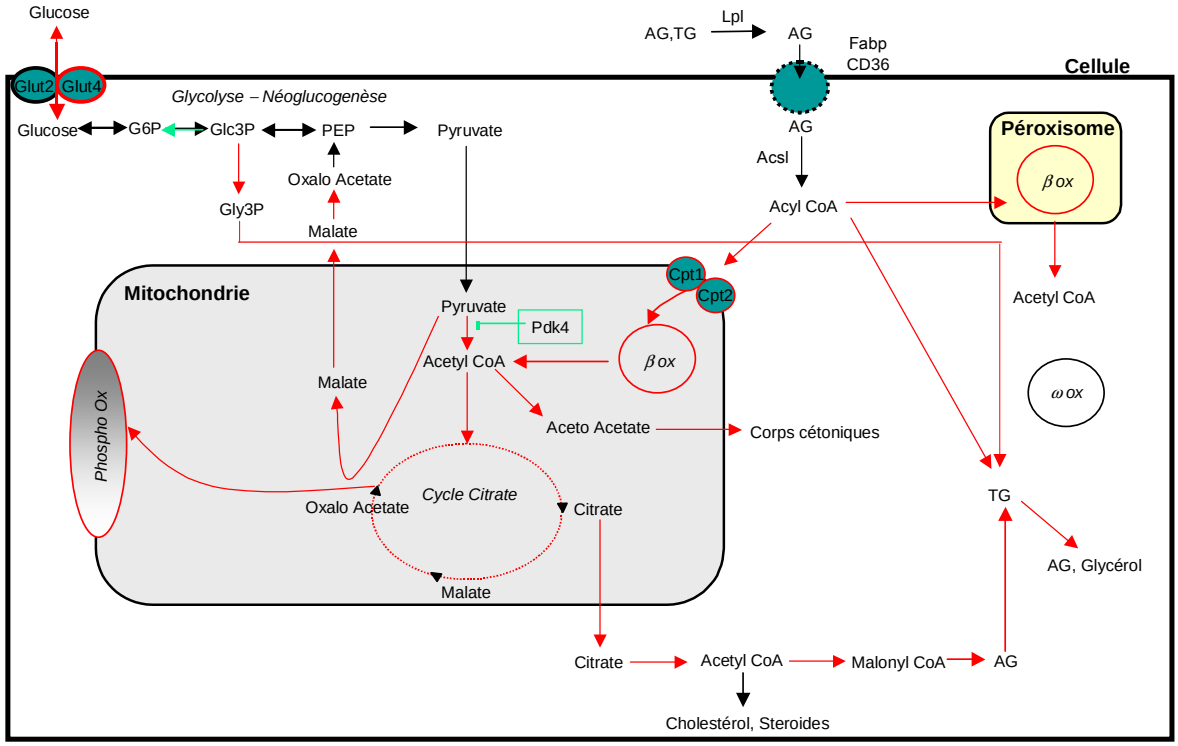

FIG. 5.9 : Modifications transcriptomiques observées dans le tissu adipeux inguinal pour les souris db/db traitées à la rosiglitazone par rapport aux souris db/db non traitées

TG : triglycérides, AG: acides gras, G6P : glucose 6-phosphate, Glc3P: glycéraldéhyde 3-phosphate, Gly3P : glycérol 3-phosphate, PEP: phosphoénolpyruvate, CoA : coenzyme A, ox : oxydation. Glut : transporteur de glucose, Lpl : lipoprotéine lipase, Fabp : protéine de liaison d'acides gras, CD36 : antigène CD36, Acsl : Acyl coenzyme A synthéthase à longue chaîne, Pdk4 : Pyruvate déshydrogénase kinase 4, Cpt : carnitine palmitoyltransférase.

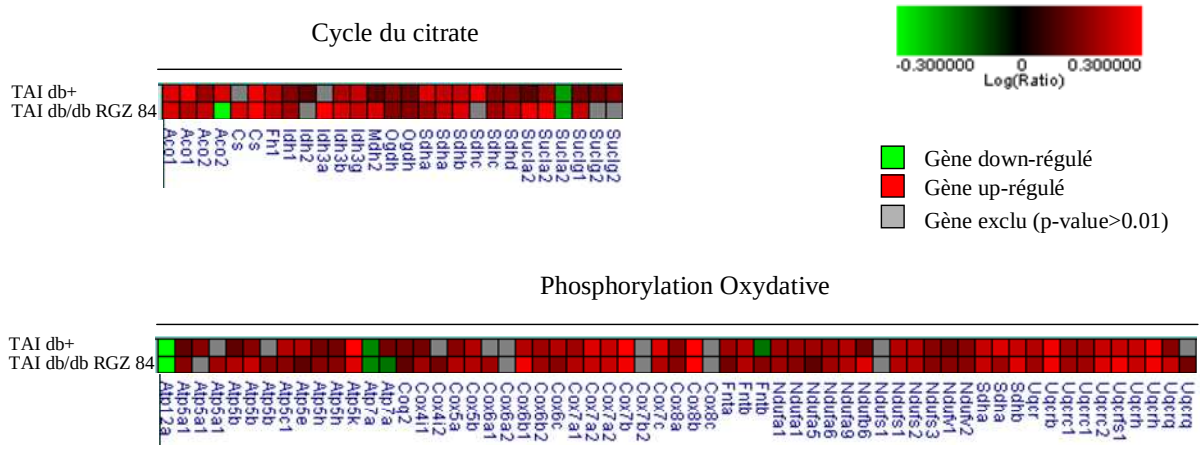

FIG. 5.10 : Visualisation des modifications transcriptomiques dans le cycle du citrate et la phosphorylation oxydative dans le TAI
Les résultats sont présentés pour les souris diabétiques db+ et les souris db/db traitées à la rosiglitazone à la dose 84 µmol/kg. La référence considérée est le groupe contrôle des souris db/db non traitées. RGZ : rosiglitazone.

5.3.3 Lien avec les séquences sélectionnées

Observations transcriptomiques et biologiques ont pu dans la plupart des cas être reliées. En effet, au niveau du foie, la rosiglitazone semble empirer la situation au niveau des graisses par rapport à des souris normales aux deux niveaux d'observations. De même, ces deux types d'informations concordent pour une normalisation dans le TAI. Nous avons dans un second temps voulu étudier la capacité des séquences sélectionnées par NSCRob à expliquer les différences biologiques observées. Ces séquences sont en effet censées expliquer au mieux les différences entre rosiglitazone et SCOMP. Pour cela, les séquences sélectionnées qui appartiennent au métabolisme ont été considérées pour la dose 84 µmol/kg. Nous nous sommes concentrés sur les voies métaboliques pouvant être interprétées directement en termes de paramètres biologiques (glycolyse, néoglucogenèse, dégradation et synthèse des triglycérides et acides gras) dans le muscle squelettique, le foie et le TAI.

5.3.3.1 Séquences sélectionnées dans le muscle squelettique

Dans le muscle squelettique (Tableau 5.6), une différence est trouvée au niveau de l'oxydation des acides gras avec Acsl4 (Acyl-Coenzyme A synthétase). Néanmoins, la différence la plus intéressante concerne Hk1 (Hexokinase). L'hexokinase régule la première étape irréversible de la glycolyse, voie de dégradation du glucose. Le gène est plus « up-régulé » avec la rosiglitazone qu'avec le SCOMP par rapport aux souris db/db non traitées.

Cette observation pourrait être liée à une diminution plus importante de la glycémie avec la rosiglitazone.

Voie métabolique	Gène(s)	RGZ	SCOMP
Glycolyse	Hk1	↑↑	↑
Oxydation des acides gras - Transport des acides gras	Acsl4	↓	↑

TAB. 5.6 : Séquences sélectionnées par NSCRob dans le muscle squelettique qui appartiennent à des voies métaboliques d'intérêt
Les flèches vers le haut et vers le bas correspondent respectivement à une « up-régulation » et une « down-régulation » par rapport aux db/db non traitées. RGZ : rosiglitazone.

5.3.3.2 Séquences sélectionnées dans le foie

Dans le foie (Tableau 5.7), les résultats sont plus difficiles à interpréter. En effet, certaines séquences d'une même voie métabolique sont plus exprimées avec le composé SCOMP et d'autres le sont plus avec la rosiglitazone. Il est donc difficile de conclure pour la β-oxydation péroxisomale et pour le transport des acides gras. De plus, le gène Ehhadh (Enoyl-CoA hydratase/3-hydroxyacyl CoA déshydrogénase) peut être impliqué dans la synthèse ou la dégradation des acides gras. Il est néanmoins possible d'émettre l'hypothèse qu'il y aurait plus de β-oxydation mitochondriale, plus d'ω-oxydation, plus de transport des lipides et moins de synthèse des triglycérides (notamment avec Agpat6, enzyme initiatrice de la synthèse des triglycérides) avec le SCOMP. Ces observations sont cohérentes pour les triglycérides, puisque leur densité est en effet moins augmentée avec le composé SCOMP qu'avec la rosiglitazone. En revanche, le SCOMP augmente plus la densité d'acides gras que la rosiglitazone dans le foie et ces résultats sont plus difficilement expliqués par les observations transcriptomiques. On peut émettre l'hypothèse qu'avec moins d'utilisation des acides gras dans la synthèse des triglycérides et plus d'import de lipides, l'augmentation de la β-oxydation mitochondriale avec le SCOMP ne suffit pas à faire baisser les quantités d'acides gras libres.

Voie métabolique	Gène(s)	RGZ	SCOMP
Oxydation des acides gras - β-oxydation mitochondriale	Acadl, Dci	↑	↑↑
Oxydation des acides gras - β-oxydation peroxisomale	Acaa1a	↑↑	↑
Oxydation des acides gras - β-oxydation peroxisomale	Ech1	↑	↑↑
Oxydation des acides gras - β-oxydation peroxisomale + Synthèse des acides gras	Ehhadh	↑	↑↑
Oxydation des acides gras - Transport des acides gras	Acsl1	↑↑	↑
Oxydation des acides gras - Transport des acides gras	Cpt1b	↑	↑↑
Oxydation des acides gras - ω-oxydation	Cyp4a10, Cyp51	→	↑
Oxydation des acides gras - ω-oxydation	Cyp4a12a, Cyp4a14	↑	↑↑
Synthèse des triglycérides	Agpat6	↑	→
Transport des lipides	Fabp1	→	↑
Transport des lipides	Fabp2	↑	↑↑

TAB. 5.7 : Séquences sélectionnées par NSCRob dans le foie qui appartiennent à des voies métaboliques d'intérêt

Les flèches vers le haut correspondent à une « up-régulation », celles vers le bas et vers la droite à une « down-régulation » et une non-régulation, par rapport aux db/db non traitées. RGZ : rosiglitazone.

5.3.3.4 Séquences sélectionnées dans le tissu adipeux inguinal

Dans le TAI (Tableau 5.8), on trouve que des gènes impliqués dans l'import d'acides gras dans la mitochondrie (la carnitine palmitoyltransférase, Cpt1b) et dans la β-oxydation mitochondriale sont plus exprimés avec la rosiglitazone. Ceci recoupe les observations biologiques qui montrent des quantités d'acides gras dans le TAI moins importantes pour la rosiglitazone que pour le SCOMP.

Voie métabolique	Gène(s)	RGZ	SCOMP
Oxydation des acides gras - β-oxydation mitochondriale	Acaa2, Dci	↑↑	↑
Oxydation des acides gras - Transport des acides gras	Cpt1b	↑↑	↑

TAB. 5.8 : Séquences sélectionnées par NSCRob dans le tissu adipeux inguinal qui appartiennent à des voies métaboliques d'intérêt
Les flèches vers le haut correspondent à une « up-régulation » par rapport aux db/db non traitées.
RGZ : rosiglitazone.

5.3.4 Conclusion

Ces observations, faites pour chaque organe au niveau transcriptomique et au niveau biologique, sont récapitulées dans la Figure 5.11. Beaucoup d'informations trouvées au niveau transcriptomique peuvent donc être directement reliées aux observations biologiques : quantité d'acides gras et β-oxydation dans le TAI, glycolyse dans le muscle et glycémie, quantité et synthèse de triglycérides dans le foie. Dans la majorité des cas, les séquences sélectionnées par NSCRob semblent donc pertinentes. Il ne faut néanmoins pas oublier que des modifications post-transcriptionnelles sont à prendre en compte pour déterminer l'activité d'une enzyme : traduction de l'ARN en protéine, phosphorylation de la protéine, ... Il est donc normal d'observer quelques désaccords entre les résultats transcriptomiques et lipidomiques.

Ces observations permettent néanmoins d'émettre au moins trois hypothèses : (i) l'augmentation des graisses dans le foie avec les traitements pourrait être due à un dérèglement des échanges de lipides avec le plasma ; (ii) la diminution des graisses dans le plasma pourrait également être liée à l'apparition d'adipocytes plus fonctionnels retrouvant leur rôle de stockage et d'oxydation des triglycérides ; (iii) la différence de glycémie entre les souris traitées à la rosiglitazone et au SCOMP pourrait être en partie associée à la différence d'expression observée pour l'hexokinase (Hk1) dans le muscle, et donc à une différence au niveau de la glycolyse. Cela ne signifie pas que seul le muscle permet de réguler la glycémie (nous avons montré que la néoglucogenèse était diminuée dans le foie), mais seulement que la différence, d'un point de vue transcriptionnel, entre les deux composés réside principalement à ce niveau.

Ces analyses des différences entre rosiglitazone et SCOMP peuvent également permettre de proposer des gènes clés à valider par une analyse en qPCR (quantitative Polymerase Chain Reaction). D'autre part, la cohérence au niveau des résultats biologiques permet d'accorder une plus grande confiance aux listes de séquences sélectionnées. Il serait alors intéressant d'analyser les gènes n'appartenant pas au métabolisme, par exemple les gènes impliqués dans l'inflammation. En effet, l'inflammation est partie prenante de l'extension du TAI observée chez les rongeurs obèses avec une infiltration de macrophages [82].

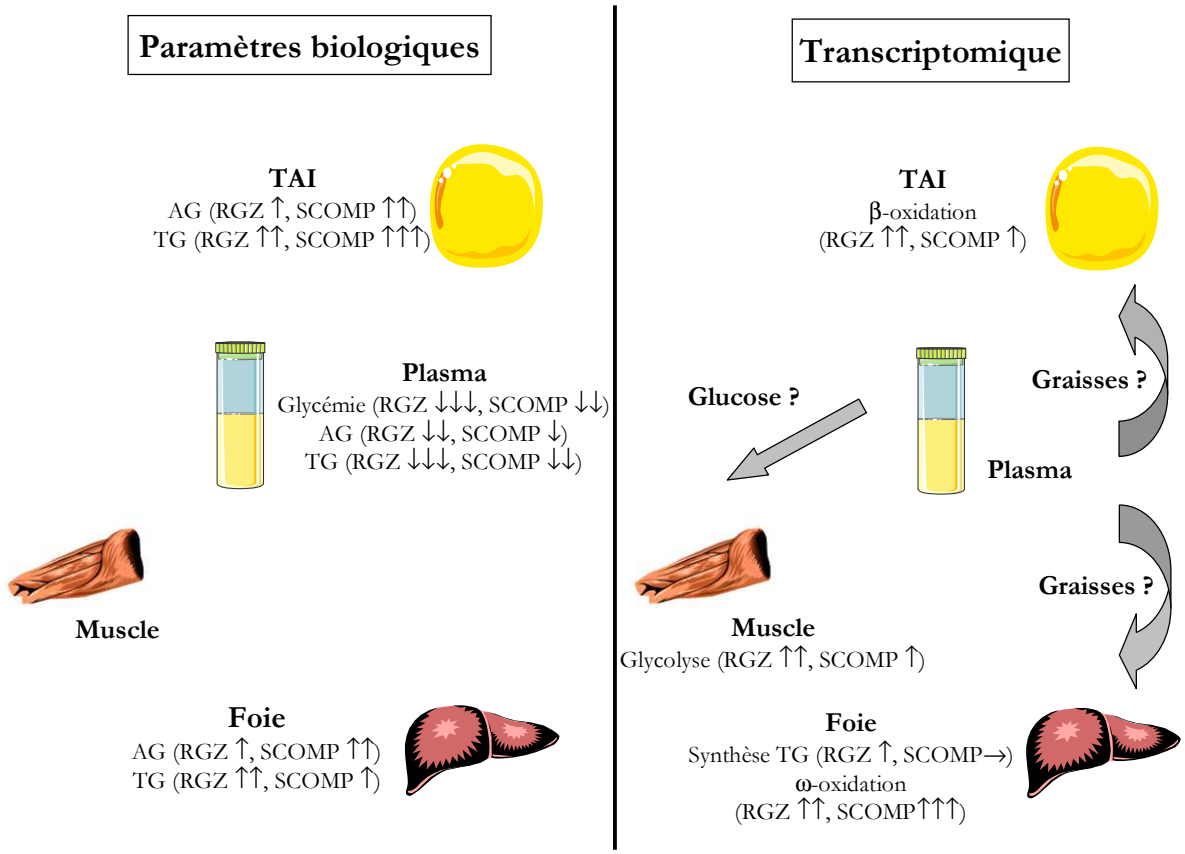

FIG. 5.11 : Lien entre paramètres biologiques et séquences sélectionnées
TAI : tissu adipeux inguinal, AG : acides gras, TG : triglycérides, RGZ : rosiglitazone
Les flèches vers le haut (respectivement vers le bas) correspondent à une augmentation (respectivement une diminution) avec un traitement par rapport aux souris diabétiques db/db non traitées.

Conclusion générale et perspectives

Problématique

Dans le cadre de la recherche de médicaments contre le diabète de type 2, l'Institut de Recherches Servier a développé un agoniste PPARγ partiel, appelé SCOMP. Afin de mieux évaluer son efficacité et son innocuité par rapport à l'offre médicamenteuse existante, ce composé a été comparé à un agoniste PPARγ de référence, la rosiglitazone, chez des souris diabétiques db/db. Des analyses en puces à ADN ont été menées sur trois tissus (muscle squelettique, foie et tissu adipeux inguinal), pour trois doses de traitements (28 µmol/kg, 84 µmol/kg et 280 µmol/kg). L'objectif de cette thèse était d'étudier les différences principales entre rosiglitazone et SCOMP, essentiellement à partir des données de puces à ADN. Ce type d'analyse fournissant des quantités très importantes de données (plus de 40000 ARNs mesurés pour chaque lame), l'exploitation des résultats a nécessité l'utilisation d'outils mathématiques et statistiques.

L'analyse de données de puces à ADN est un problème complexe et bien connu en statistique. La technologie des puces à ADN permet de mesurer le niveau d'expression de plusieurs milliers de gènes dans un échantillon biologique. Le nombre d'observations disponibles est cependant souvent limité à quelques dizaines. Ce déséquilibre entre nombres d'observations et nombres de variables est appelé « fléau de la dimension » et sort du cadre classique de la statistique. L'étude des différences entre rosiglitazone et SCOMP a nécessité l'application de méthodes de sélection de variables pour cibler les gènes d'intérêt. Plusieurs types de méthodes ont déjà été appliqués à des données de puces à ADN. On peut citer parmi celles-ci l'analyse en composantes principales, les arbres de classification, ou encore les algorithmes génétiques, … Suite à une étude bibliographique, nous avons sélectionné parmi celles-ci trois méthodes à tester : T-test, Nearest Shrunken Centroids (NSC) et Support Vector Machines – Recursive Feature Elimination (SVM-RFE). Cependant, il peut parfois être difficile de choisir un nombre de séquences optimal avec ces méthodes et l'impact de la variabilité technique des puces à ADN sur leurs résultats reste mal connu.

Les travaux de cette thèse ont consisté à comparer les trois méthodes de sélection de variables précitées et à mettre au point une nouvelle méthodologie, MetRob, permettant de répondre aux questions de choix d'un nombre de séquences optimal et de prise en compte de l'impact de la variabilité technique. La méthodologie que nous avons développée permet de construire des listes de séquences discriminantes robustes et reproductibles. La robustesse a été évaluée comme la capacité d'une méthode de sélection de variables à fournir une liste de séquences

stable sous l'effet de perturbations des données. Cette évaluation a requis la génération de jeux de données virtuels reproduisant la variabilité technique des puces à ADN, car la réalisation de réplicats techniques en quantité suffisante n'était économiquement pas faisable. Comme les modes de perturbation classiques n'étaient pas adaptés à nos données, une étude de variabilité technique à petite échelle a tout d'abord été menée afin de pouvoir reproduire un bruit réaliste. Le problème du choix du nombre de séquences et le manque de prise en compte de la robustesse ont été gérés conjointement en choisissant le nombre de séquences maximisant la robustesse

Résultats

La méthodologie développée a été testée sur des jeux de données simulées et sur des jeux de données réelles. Les données simulées ont été obtenues à partir du logiciel SIMAGE, qui permet de reproduire les caractéristiques de données de puces à ADN. Des variables discriminantes en quantité et en amplitude connues ont ensuite été introduites. Les données réelles sont issues de la comparaison entre la rosiglitazone et le SCOMP.

Les résultats sur données simulées ont montré que les méthodes T-test et NSC étaient plus robustes que la SVM-RFE et que les séquences sélectionnées possédaient un bon pouvoir discriminant quelque soit la méthode de sélection de variables utilisée. Bien que l'on observe une diminution du nombre de séquences sélectionnées et une sensibilité (capacité à détecter les variables discriminantes) parfois très faible avec MetRob, la valeur prédictive positive (confiance accordée aux prédictions) est grandement améliorée. L'intérêt principal de MetRob réside donc dans l'obtention de listes de séquences fiables. Sur données réelles, les séquences sélectionnées possèdent également un bon pouvoir discriminant. La méthode de SVM-RFE obtient toujours les résultats les moins robustes et c'est la méthode NSC qui semble la plus intéressante. C'est donc elle qui a été choisie pour une analyse biologique des résultats. Par ailleurs, une étude de l'impact du nombre d'observations a été initiée. Sur données réelles, on a pu observer qu'augmenter le nombre d'observations réduisait le nombre de séquences sans en ajouter de nouvelles, avec la méthode NSC. Cette remarque est également vraie sur données simulées pour le jeu de données se rapprochant le plus des données réelles.

Nous avons enfin cherché à valider les résultats obtenus par l'association de MetRob et NSC avec une analyse biologique des listes de séquences. Dans un premier temps, nous avons considéré les enrichissements en séquences des voies métaboliques. 18 voies métaboliques d'intérêt ont été choisies et nous avons cherché à déterminer si les séquences de ces voies étaient préférentiellement sélectionnées par MetRob. Les résultats obtenus sont cohérents avec les niveaux d'expression de PPAR dans les tissus et avec les connaissances actuelles du rôle de PPAR dans le contrôle du métabolisme. Cette analyse a également permis de renforcer l'hypothèse que les séquences les plus intéressantes sont conservées quand on augmente le nombre d'observations. Nous avons ainsi pu montrer que les gènes sélectionnés par l'association de MetRob et NSC étaient biologiquement pertinents. La nouvelle méthodologie

que nous avons développée permet donc de sélectionner des séquences pertinentes à la fois du point de vue de la discrimination et du point de vue biologique. Dans un second temps, un parallèle a été établi entre les observations faites au niveau transcriptomique et les variations de différents paramètres biologiques (glycémie, acides gras, triglycérides) pour la dose 84 µmol/kg. Cette étude nous a permis d'émettre plusieurs hypothèses quant à l'effet des traitements. L'augmentation des graisses dans le foie et leur diminution dans le plasma pourraient être liées à un dérèglement des échanges de lipides entre foie et plasma et à l'apparition d'adipocytes plus fonctionnels. De plus, la différence de glycémie observée entre les souris traitées à la rosiglitazone et celles traitées au SCOMP est potentiellement associée à une différence au niveau de la glycolyse musculaire.

La méthodologie mise au point au cours de cette thèse permet donc d'obtenir des résultats pertinents, tant du point de vue de la discrimination, que du point de vue biologique. L'outil développé a été mis à la disposition de l'Institut de Recherches Servier. Son application à d'autres problématiques biologiques pourrait permettre, d'une manière plus générale, une aide à l'interprétation pharmacologique de l'étude des différences entre deux composés.

Positionnement par rapport aux travaux récents dans le domaine

Nous avons tenté de replacer ce travail dans le contexte des travaux récents de sélection de gènes différentiellement régulés, à partir de données de puces à ADN.

De nouvelles démarches proches des méthodes testées au cours de cette thèse ont été récemment développées. On peut citer parmi elles la SVM-RNE (Support Vector Machine – Recursive Network Elimination) et BAHSIC (Backward Elimination and Hilbert-Schmidt Independance Criterion). La SVM-RNE est une adaptation de la SVM-RFE qui est présentée dans ce manuscrit [83]. Au lieu de supprimer un gène à chaque itération, ce sont des groupes de gènes qui sont éliminés. Ces groupes sont définis comme des modules de gènes au sein d'un réseau. Une telle approche permet une interprétation biologique des résultats plus facile. Elle nécessite cependant que les gènes soient présents dans la base de données d'interactions utilisée. BAHSIC est un cadre général permettant de réaliser des analyses de classification multiclasse et de régression. Il regroupe plusieurs méthodes existantes, parmi lesquelles le T-test et NSC [84]. Concrètement, il s'agit de méthodes d'éliminations successives des variables utilisant des noyaux. Les auteurs concluent à une meilleure efficacité des noyaux linéaires dans les cas généraux, ce qui est cohérent avec les résultats que nous avons pu obtenir lors du choix de noyau pour les SVMs.

L'analyse de données de puces à ADN tend aujourd'hui à inclure des connaissances biologiques provenant d'autres sources d'information. Par exemple, Jeffery et al. [85] recherchent des différences d'activités de facteurs de transcription en associant des informations sur les motifs des séquences promotrices des gènes à des données transcriptomiques. Phan et al. [86] déterminent quant à eux quelle est la méthode de sélection

de variables adaptée à leur jeu de données à partir de connaissances biologiques a priori. Une méthode est considérée comme efficace si elle accorde un bon classement à des biomarqueurs connus. Ce type de démarches est très intéressant et représente certainement l'avenir des analyses de données biologiques. Les connaissances supplémentaires à leur utilisation ne sont néanmoins pas toujours disponibles et il reste utile de disposer d'approches générales travaillant uniquement à partir des données de puces à ADN.

Dans le même esprit d'intégration, la combinaison de différentes méthodes est également beaucoup utilisée afin d'améliorer la fiabilité et la robustesse des résultats. ArrayMining est une application web qui permet de sélectionner des gènes à partir des classements obtenus par différentes méthodes [87]. Cependant, le problème du nombre de gènes à sélectionner demeure car il incombe à l'utilisateur de le fixer. L'algorithme présenté dans [88] cherche à obtenir un nombre très limité de gènes réellement discriminants. Il s'agit essentiellement d'une superposition de méthodes de sélection de variables. La robustesse de ces méthodes est essentiellement évaluée comme la stabilité des résultats par ajout ou suppression d'échantillons, et non par perturbation des valeurs des variables comme nous l'avons fait. Ce type d'approche s'apparenterait plus à notre étude de l'impact du nombre d'observations. Des techniques de sélection de variables d'ensemble sont présentées dans [89] et évaluées à partir de cette définition de la robustesse. L'application d'une méthode sur différents échantillons bootstrap permet la génération de plusieurs listes de gènes, qui sont ensuite combinées. Toutes ces approches de combinaisons de méthodes sont à rapprocher de l'une des perspectives proposées ci-dessous.

Perspectives

Les problématiques abordées pendant ce travail de thèse nous permettent d'identifier les perspectives futures de cette étude selon les axes suivants :

- Nous avons évoqué que l'utilisation de MetRob permettait d'obtenir de bonnes valeurs prédictives positives, mais des sensibilités faibles. Le couplage de différentes méthodes de sélection de variables pourrait permettre de compenser ce problème. En effet, nous avons pu observer que les séquences sélectionnées par la méthode de SVM-RFE lui étaient parfois spécifiques. Combiner des méthodes de sélection de variables avec des concepts très différents pourrait alors permettre d'améliorer la sensibilité tout en conservant une bonne valeur prédictive positive.

- L'étude de l'impact du nombre d'observations sur les résultats a seulement été initiée. Il serait intéressant, à plus long terme, de poursuivre cette étude avec des nombres d'observations plus élevés, et surtout en considérant des nombre d'observations intermédiaires, de manière à pouvoir choisir un optimum.

- Nous avons ici étudié les séquences sélectionnées du point de vue du métabolisme. Du fait de la nature de la pathologie étudiée et de la cible thérapeutique considérée, cette approche était la plus évidente pour une validation des résultats de MetRob. Maintenant que la confiance accordée aux résultats a été renforcée, il serait intéressant d'étudier plus précisément les rôles des séquences sélectionnées qui ne sont pas impliquées dans le métabolisme. Une telle étude pourrait permettre d'émettre de nouvelles hypothèses sur le mode d'action des agonistes PPAR.

- Dans le cadre de cette étude, nous avons comparé les deux agonistes PPAR à différentes doses. Comme la méthodologie développée se voulait générale et utilisable dans le cadre d'une comparaison simple entre deux classes, les relations entre les doses n'ont pas été étudiées.. Les séquences sélectionnées représentent donc les différences les plus importantes entre rosiglitazone et SCOMP à une dose fixée. Une comparaison directe entre les listes de séquences associées aux trois doses est néanmoins difficile. Les effets dose-dépendants n'ont aucune raison d'être linéaires ou même d'évoluer en parallèle pour les deux composés. Les différences entre rosiglitazone et SCOMP peuvent donc se situer à des niveaux différents en fonction de la dose. D'autres approches peuvent néanmoins être envisagées pour étudier les données dans leur globalité : comparaison des deux produits à des doses différentes, analyse de variance sur tous les jeux de données, regroupement des données par produit, puis étude des différences toutes doses confondues, …

Annexe A :

Précisions d'ordre biologique

A.1 Caractérisation in vitro du composé SCOMP

Le produit SCOMP a été synthétisé à l'Institut de Recherches Servier et est un modulateur sélectif de PPARγ (SPPARM) qui est également capable d'activer PPARα. Son activité PPARγ a été mesurée par transfection transitoire dans des lignées cellulaires 3T3L1 (tissu adipeux de souris) et HepG2 (foie humain). Le système rapporteur utilisé consiste en un récepteur Gal4 modifié : le brin d'ADN est composé du domaine de liaison au ligand PPARγ humain et d'une région codant pour le domaine de liaison à l'ADN de la protéine Gal4 (Figure A.1). Ces récepteurs activés par l'agoniste SCOMP stimulent la production de luciférase qui est détectée

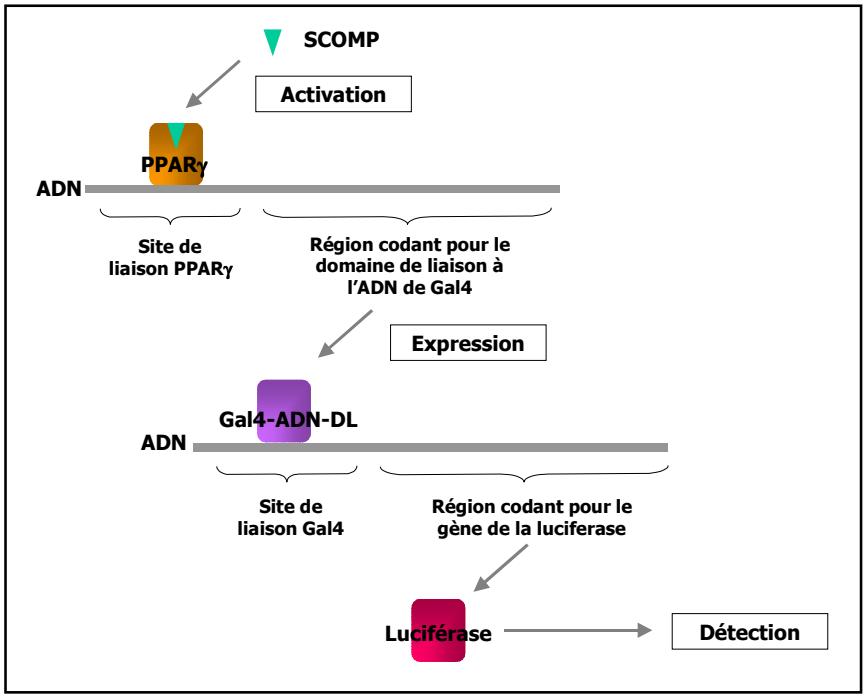

FIG. A.1 : Transfection transitoire avec un système rapporteur Gal4

Les résultats sont présentés dans le Tableau A.1. Emax est l'activation maximale du site de liaison PPARγ par le composé (la rosiglitazone représentant la référence avec une activation à 100%). EC_{50} est la concentration du composé pour laquelle 50% de la réponse Emax est atteinte. Avec un Emax de 90% dans les cellules HepG2, le SCOMP a une bonne affinité pour PPARγ dans le foie. Dans les cellules 3T3L1, son activation est plus faible avec un Emax de 20% et un EC50 élevé de 3.26 µM. Le SCOMP a donc essentiellement des propriétés de SPPARM dans le tissu adipeux. Il faut également noter qu'à concentration élevée (10µM), le SCOMP active le PPARα humain dans la lignée cellulaire Cos7 : il induit 141% de la réponse spécifique PPARα du WY14,643 (activation de 100% pour 50µM).

	3T3L1		**HepG2**	
	Emax (%)	**EC50 (µM)**	**Emax (%)**	**EC50 (µM)**
Rosiglitazone	100	0.41	100	0.25
SCOMP	20	3.26	90	0.44

TAB. A.1 : Activation de PPARγ par la rosiglitazone et le SCOMP dans les lignées cellulaires 3T3L1 et HepG2
3T3L1 : tissu adipeux de souris, HepG2 : foie humain.

A.2 Détail du protocole de marquage et d'amplification de l'ARN

Avant de pouvoir être hybridé sur des puces à ADN, l'ARN messager d'un échantillon biologique doit être amplifié et marqué par des fluorochromes pour être détectable. Le protocole utilisé est le protocole « Agilent Low RNA Input Linear Amplification Kit » détaillé sur la Figure A.2 [90]. L'ARN sens est extrait de l'échantillon biologique. Cet ARN est incubé en présence d'un promoteur T7 (promoteur associé à une queue poly-T) et d'une transcriptase inverse, la MMLV-RT (Moloney Murine Leukemia Virus Reverse Transcriptase). L'enzyme induit la synthèse des ADN double-brins correspondants à l'ARN initial. Ces brins d'ADN sont chauffés pour être séparés puis mis en présence d'ARN polymérase T7 et de ribonucléotides dont certains cytidines sont marquées par un fluorochrome (Cy3 ou Cy5 selon l'échantillon). L'ARN polymerase T7 reconnaît le promoteur antisens qui est sur le brin d'ADN sens et induit la synthèse d'une centaine de brins d'ARN antisens qui sont ainsi amplifiés et marqués. Ce sont ces brins d'ARN antisens marqués qui seront hybridés sur les puces à ADN.

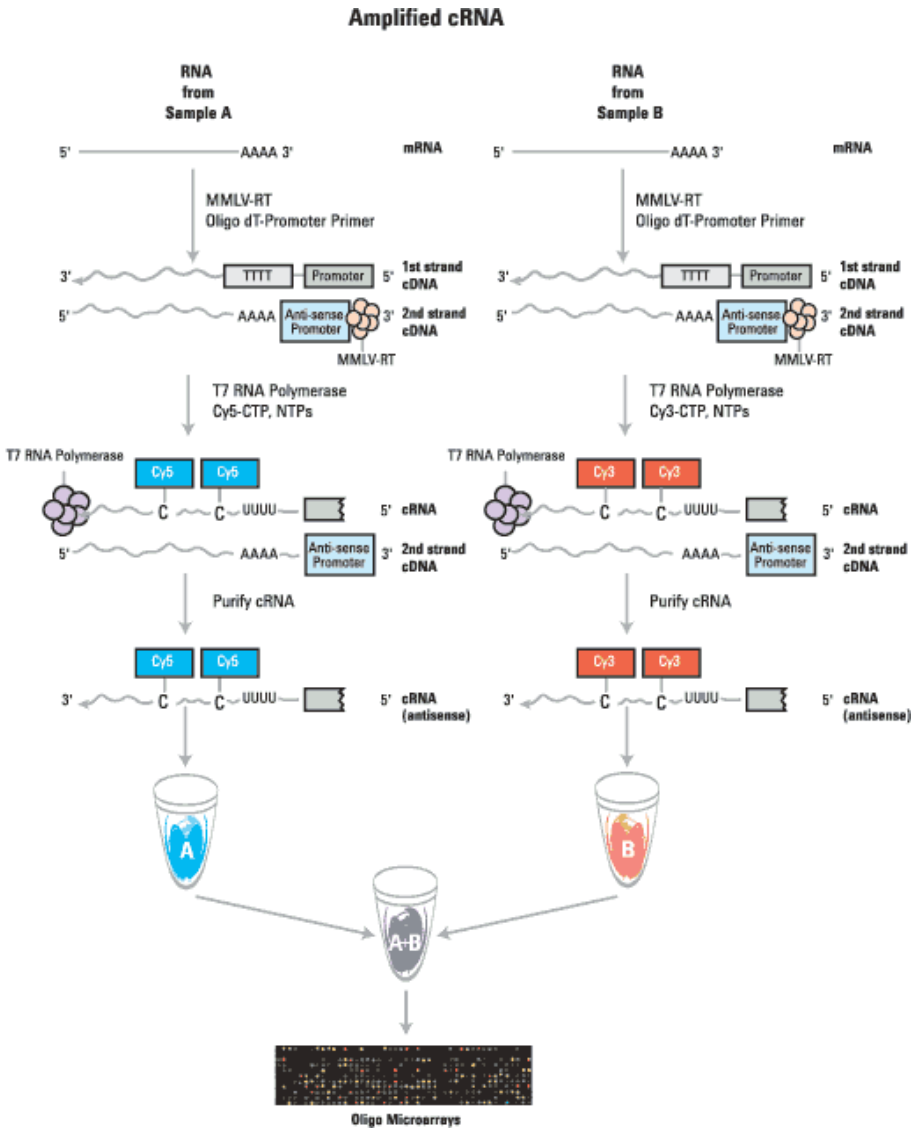

FIG. A.2 : Protocole Agilent de marquage et d'amplification de l'ARN [90]

A.3 Modèle ob/ob et protocole expérimental

D'autres molécules agonistes PPAR ont également été testées sur le modèle de souris ob/ob mais les résultats obtenus ont essentiellement été utilisés pour paramétrer les différentes méthodes mathématiques mises en place. Ces agonistes sont le fénofibrate (agoniste α) et le muraglitazar (agoniste mixte α/γ) qui ont déjà été cités dans le corps du manuscrit, ainsi que trois agonistes mixtes α/γ : le farglitazar, le tesaglitazar et un produit Servier, le SPROD.

Les souris ob/ob [15] présentent une mutation sur le gène codant pour la leptine. La leptine, hormone régulant la prise de nourriture chez la souris, n'est alors pas exprimée. Ces souris sont obèses, hyperphagiques et développent un syndrome similaire au diabète de type 2 : hyperglycémie, intolérance au glucose et hyperinsulinémie. L'hyperinsulinémie est probablement le résultat de la prise de poids et l'insulino-résistance est associée à une surproduction hépatique de glucose. Au niveau moléculaire, l'insulino-résistance semble être entre autres due à une diminution de la fixation de l'insuline à ses récepteurs.

Deux expériences ont été réalisées en juin 2004 et en avril 2006 sur des souris ob/ob (Figure A.3). La première expérience comportait un groupe témoin et quatre groupes de traitements : fénofibrate (350 mg/kg), rosiglitazone (10 mg/kg), tesaglitazar (3 mg/kg) et SPROD (3 mg/kg). La seconde expérience comportait un groupe témoin et quatre groupes de traitements avec trois agonistes mixtes PPARα/γ (tesaglitazar : 3 mg/kg, farglitazar : 3 mg/kg et muraglitazar : 3 mg/kg) et un agoniste SPPARM (SCOMP : 3 mg/kg). Chaque groupe était constitué de 6 animaux. Les souris ont été traitées durant 10 jours puis sacrifiées. Le foie, le cœur et le tissu adipeux inguinal (TAI) de ces souris ont été prélevés pour être analysés en puces à ADN. Les puces à ADN utilisées étaient des lames de souris Agilent 22k plus anciennes que les lames 4*44k utilisées pour les expériences chez les souris db/db. Des mesures biologiques ont également été réalisées durant l'expérience : poids des animaux, glycémie et taux d'hémoglobine glycquée (HbA1c).

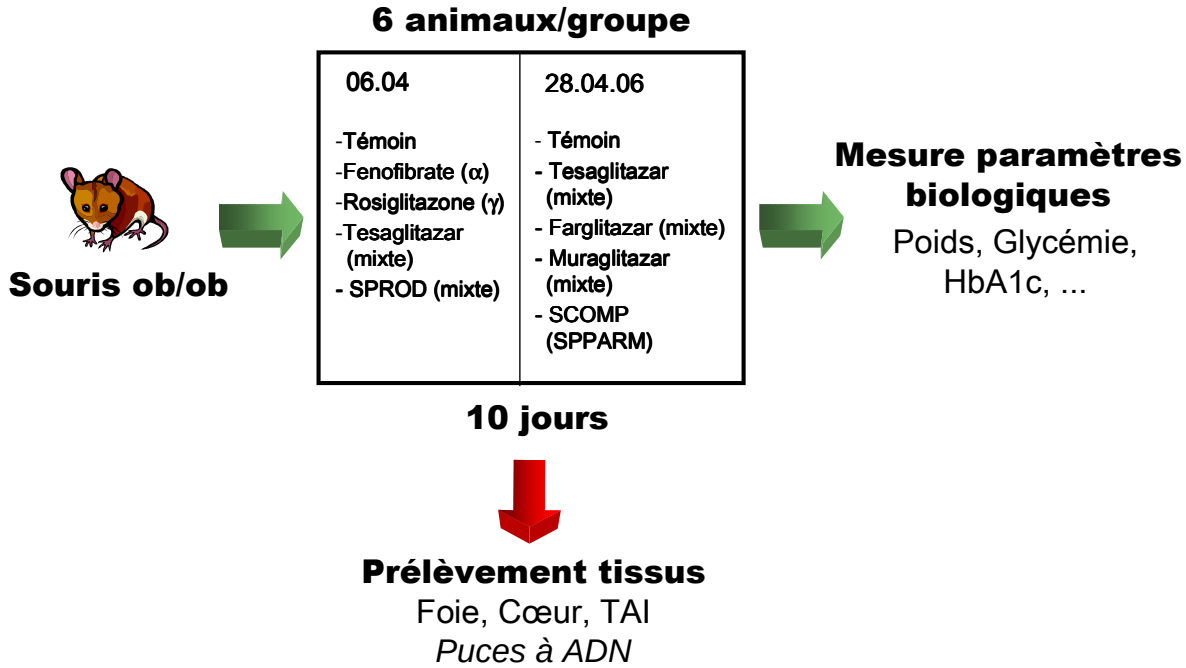

FIG. A.3 : Protocole expérimental pour l'étude des agonistes PPAR chez les souris ob/ob
TAI = Tissu adipeux inguinal, HbA1c = Hémoglobine glycquée

A.4 Paramètres biologiques et analyse lipidomique

Lors des expériences PPAR, la glycémie a été mesurée à partir de prélèvements sanguins. En parallèle une analyse en lipidomique a été menée pour le plasma, le foie et le TAI. Les quantités de 11 acides gras et de 4 triglycérides ont été mesurées par UPLC-MS (Ultra Performance Liquid Chromatography-tandem Mass Spectrometric). Elles ont été données sous forme de concentrations pour le plasma et elles ont été normalisées par le poids de l'organe pour le foie et le TAI. Au final, trois acides gras (acides palmitique, linoléique et oléique) et deux triglycérides (trilinoléate et trioléate) majoritaires ont été conservés. Les chiffres sont présentés dans le Tableau A.2.

	db+	**db/db**		
Paramètres (unités)	**Non traitées**	**Non traitées**	**RGZ 84 µmol/kg**	**SCOMP 84 µmol/kg**
Poids du foie (g)	1.42 ± 0.05	1.75 ± 0.03	2.85 ± 0.2	3.22 ± 0.25
Poids du TAI (g)	0.17 ± 0.02	2.08 ± 0.08	2.1 ± 0.13	1.86 ± 0.08
Glycémie (g/l)	5.2 ± 0.5	16.4 ± 1	5.5 ± 0.7	8.8 ± 2.1
TG plasmatiques (nmol/ml)				
Trilinoleate	26.8 ± 10.2	55 ± 5	10.2 ± 3	20.4 ± 3.4
Trioleate	11.8 ± 3.4	20.5 ± 1	3.2 ± 0.9	9.1 ± 3.2
AG plasmatiques (nmol/ml)				
Acide palmitique	79.5 ± 8.7	92.2 ± 4	52.6 ± 6.3	75.6 ± 6.1
Acide linoléique	64.8 ± 4	81.6 ± 3.4	44.3 ± 7.6	69.9 ± 7.2
Acide oléique	62.4 ± 8.3	87.5 ± 9	40.1 ± 8.9	69.1 ± 8.5
TG hépatiques (nmol/g)				
Trilinoleate	20 ± 3.9	142 ± 21.1	43 ± 8.9	35 ± 10.7
Trioleate	58.8 ± 12.8	871 ± 287	7746 ± 1603	4546 ± 1352
AG hépatiques (nmol/g)				
Acide palmitique	366 ± 15	463 ± 44	573 ± 47	667 ± 87
Acide linoléique	209 ± 11	296 ± 26	253 ± 17	296 ± 35
Acide oléique	186 ± 18	447 ± 78	997 ± 82	1155 ± 161
TG dans le TAI (nmol/g)				
Trilinoleate	15524 ± 4320	5131 ± 2346	5242 ± 1542	7580 ± 1043
Trioleate	16021 ± 5333	7385 ± 1961	15382 ± 5458	42456 ± 7113
AG dans le TAI (nmol/g)				
Acide palmitique	424 ± 92	263 ± 53	302 ± 67	487 ± 71
Acide linoléique	558 ± 82	362 ± 37	454 ± 97	586 ± 56
Acide oléique	65 ± 10	312 ± 31	416 ± 110	534 ± 62

TAB. A.2 : Dosage de la glycémie et résultats de l'analyse lipidomique
Les données sont fournies à ± SEM (« standard error on the mean »). RGZ : rosiglitazone.

Annexe B :

Détails sur les logiciels de traitement des données

B.1 Détail du protocole de Feature Extraction 9.5

B.1.1 Protocole général

Le protocole GE2-v5_95_Feb07 a été utilisé. C'est le protocole par défaut recommandé par Agilent pour les lames 4-44k utilisées ici. Il comporte 8 étapes (Figure B.1) qui sont détaillées dans les parties suivantes [28][91].

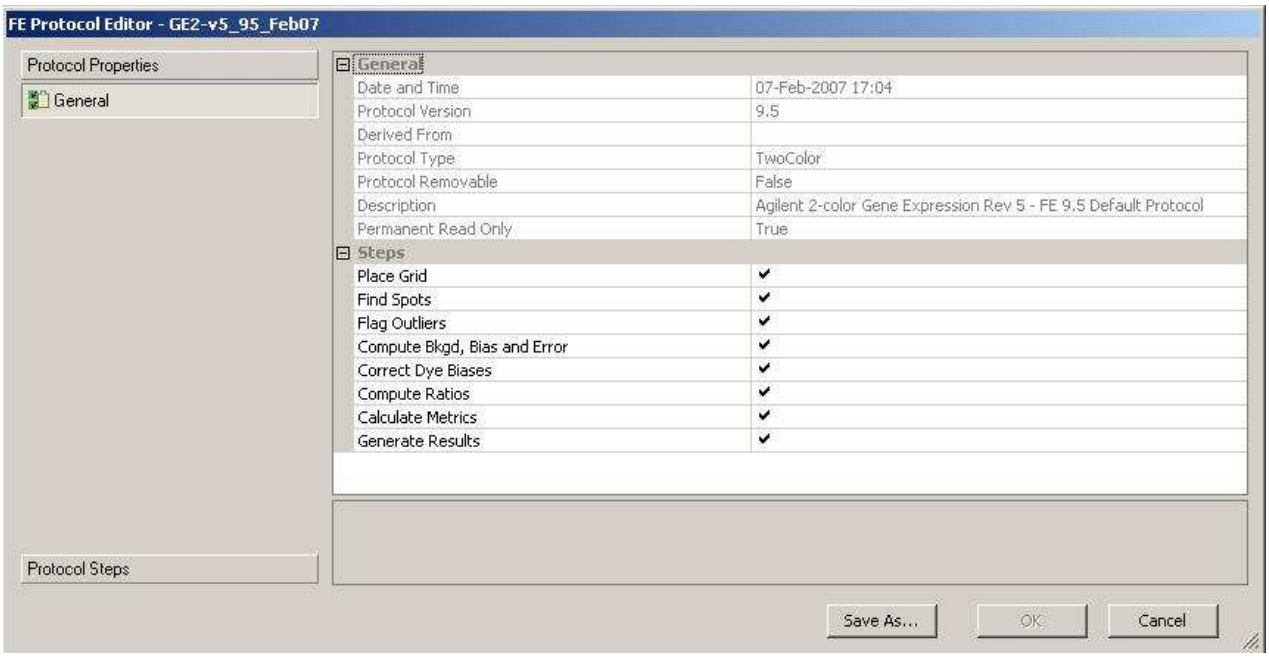

FIG. B.1 : Copie d'écran du protocole général

B.1.2 Positionnement de la grille (Place Grid)

Les spots de la lame sont localisés par le positionnement d'une grille. La méthode de positionnement utilisée est « Automatically determined » : le format de la grille est déterminé automatiquement par le logiciel. Une variance est autorisée dans le positionnement des spots et la grille peut être penchée.

B.1.3 Localisation des spots (Find Spots)

L'option par défaut est « Automatically Determine ». Dans les faits, les spots et leur taille sont identifiés, puis le signal et le bruit de fond sont définis (Spot format) et certains pixels sont rejetés avant le calcul du signal moyen (Pixel outlier rejection)

Spot Format

Le diamètre nominal de chaque spot provient du modèle de grille utilisée (dépendant des lames. Un spot a une déviation maximale autorisée par rapport à sa position sur la grille de 1.5 fois son rayon nominal. En cas de problèmes de reconnaissance de spot, cette déviation peut être augmentée manuellement.

Les notions de signal (ou feature) et de bruit de fond sur chaque spot sont définies pour les lames Agilent par une méthode appelée Cookie Cutter (Figure B.2). Le spot est représenté par le cercle noir. Son diamètre est celui donné par un fichier de modèle de lame. On définit le feature (disque vert) par son diamètre représentant 65% de celui du cercle noir. La zone d'exclusion est située entre le feature et le cercle bleu. Cette zone n'est pas utilisée dans les calculs ultérieurs. Le bruit de fond est défini par la couronne située entre le cercle bleu et le cercle rouge. Le cercle bleu a un diamètre qui représente 120% de celui du cercle noir (option exclusion zone percentage). Le cercle rouge est estimé par le logiciel.

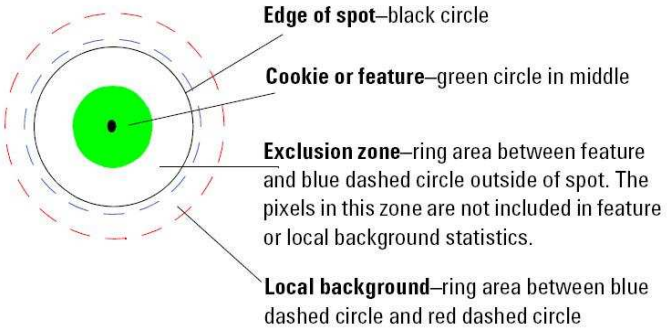

FIG. B.2 : Méthode de Cookie Cutter

Pixel outlier rejection

Il s'agit du rejet de certains pixels qui ne seront pas pris en compte dans le calcul du feature et du bruit de fond. L'option utilisée ici est le rejet des pixels basé sur l'IQR (Interquartile Range, Figure B.3). Sont rejetés les pixels pour lesquels $I > \text{mediane}(I) + (IQR/2 + \alpha \times IQR)$ ou $I < \text{mediane}(I) - (IQR/2 + \alpha \times IQR)$ avec :

- I, intensité mesurée
- mediane(I), médiane des intensités mesurées dans la région considérée (valeur centrale d'intensité telle que la moitié des pixels ont une mesure d'intensité plus petite).
- IQR, longueur de l'intervalle [*a*,*b*] tel que ¼ des valeurs observées a une intensité plus petite que a (i.e. a est le premier quartile ou 25ème percentile) et ¼ des valeurs observées a une intensité plus grande que b (i.e. b est le troisième quartile ou 75ème percentile)
- α, coefficient choisi par l'utilisateur (un coefficient pour le feature et un coefficient pour le bruit de fond) : il vaut par défaut 1.42 (probabilité de 1% qu'un pixel ne soit pas dans l'intervalle si l'intensité suit une loi normale)

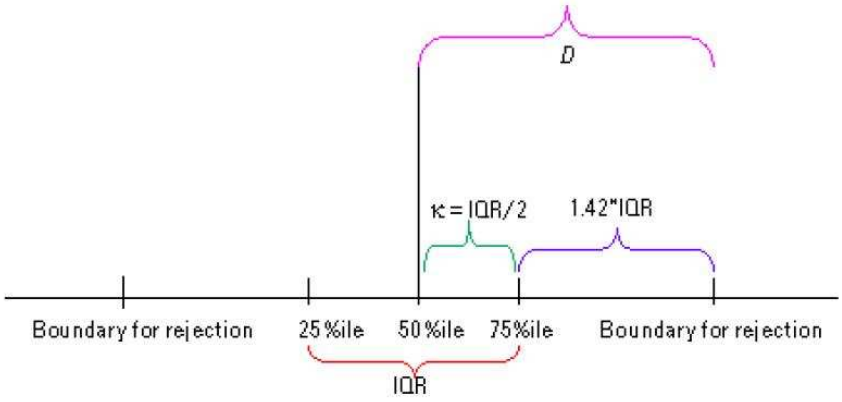

FIG. B.3 : Interquantile Range

Si plus de 50% des pixels restants dans le feature après cet algorithme sont saturés (I supérieur à environ 630000), le feature est marqué comme saturé pour la couleur correspondante. Un spot saturé pour le canal Cy3 et le canal Cy5 sera signalé dans Resolver, mais la saturation d'un seul canal ne sera pas reportée.

Statistical method for Spot Values from Pixels

Plusieurs statistiques sont calculées : moyenne, écart type, médiane, nombre de pixels et corrélations des pixels rouges et verts. C'est la moyenne qui sera utilisée dans ce protocole pour définir l'intensité.

B.1.4 Marquage des spots anormaux (Flag Outliers)

Non Uniform Outliers

Cet algorithme permet de marquer les features et les bruits de fond pour lesquels l'intensité n'est pas uniforme. Il est basé sur un modèle du bruit tel que la variabilité estimée de l'intensité des pixels est $\sigma_E^2 = Ax^2 + Bx + C$ avec :

- x, intensité moyenne des pixels pour un feature ou un bruit de fond donné
- $Ax^2 = \sigma_A^2$, sources de variabilité proportionnelles au signal (liées au marquage, à la synthèse, …)
- $Bx = \sigma_B^2$, sources de variabilité proportionnelles à la racine carrée du signal
- $C = \sigma_C^2$, sources de variabilité indépendantes du signal (liées à l'électronique du scanner, au verre, …)

Sont signalés les spots pour lesquels la variabilité réelle σ_M^2 vérifie $\sigma_M^2 > (\sigma_E^2 \times CI)$ où CI est la borne supérieure de l'intervalle de confiance obtenu à partir de la loi du chi2. La valeur de A a été estimée par des essais normal vs normal sur des puces Agilent. Les valeurs de B et C sont estimées automatiquement.

Population outlier

Cet algorithme permet de marquer les features et les bruits de fond dont les moyennes des intensités des pixels sont extrêmes au sein d'une population de réplicats sur la puce.

Si un feature a un nombre minimal de réplicat de 10, la méthode utilisée est l'IQR déjà précité avec un IQR limite de 1.42. S'il y a entre 3 et 10 réplicats, un Q-test est utilisé. C'est-à-dire que l'on calcule $Q_i = \frac{|X_i - X_{nearest}|}{|X_{max} - X_{min}|}$ avec :

- Xi intensité de la séquence considérée
- Xnearest, intensité la plus proche dans les réplicat
- Xmin, intensité minimale
- Xmax intensité maximale

Qi est comparé à un Qcritique qui dépend du nombre de réplicats. Ce Qcritique est choisi pour correspondre à un intervalle de confiance de 95%.

Les spots trouvés par cet algorithme ne sont néanmoins pas signalés dans Resolver.

B.1.5 Calcul du bruit de fond, du biais et de l'erreur

Cette étape comprend la soustraction du bruit de fond, la correction d'un effet spatial, différents tests de significativité du signal et le choix du modèle de calcul d'erreur.

Background substraction method

L'option par défaut est l'absence de soustraction du bruit de fond.

Spatial Detrend

Cet algorithme permet de corriger un effet spatial additif des données pour chaque canal d'intensité. Un ensemble de features de très faible intensité uniformément répartis sur la lame est sélectionné. Une régression loess en deux dimensions est réalisée afin d'évaluer une intensité de bruit de fond dépendante de l'espace pour chaque canal (Cy3 et Cy5). Cette intensité est ensuite soustraite au signal. Loess est l'acronyme de "Locally weighted scatterplot smoothing". Il s'agit d'une régression locale par moindres carrés pondérés. Les fonctions utilisées sont linéaires ou quadratiques. Dans le cas d'utilisation de fonctions uniquement linéaires, on parle de régression Lowess.

Modèle d'erreur

Pour les calculs d'erreurs sur le signal, la plus grande des deux erreurs suivantes est utilisée pour chaque feature.

- **Propagated error based on pixel-level statistics** : La dispersion des observations est estimée à partir de la dispersion des pixels. Ce modèle est adapté au bruit présent pour des intensités faibles (dû aux instruments et aux signaux non uniformes). Il ne capture pas les bruits présents à des intensités élevées (variabilité biologique et chimique : marquage, hybridation, lavage). Il conduit donc à sous-estimer les dispersions pour des intensités élevées.

- **Use Universal error Model** : La dispersion des observations est estimée à partir d'un modèle supposant un effet additif et un effet multiplicative ($\sigma^2=A^2+M^2I^2$). La valeur par défaut de M^2 a été estimée à partir d'expériences réalisées par Agilent (0.1 par défaut pour les deux canaux). La valeur de A^2 est estimée automatiquement. Ce modèle conduit à sous-estimer les dispersions pour de faibles intensités.

En pratique le modèle le plus utilisé est le Universal Error Model.

Feat Significance

- **IsPosAndSignif** : Cette étape teste la significativité du signal par rapport au bruit de fond qui lui est soustrait (ici le résultat du Spatial Detrend). On teste l'hypothèse nulle H_0 : Xi=0, où Xi est le signal après spatial detrend. Le test, effectué sur les moyennes des pixels, est un test de Student (basé sur l'approximation d'une beta incomplète). On suppose connue l'erreur associée à ce signal : on utilise l'erreur additive du modèle unversel divisée par 2.6. Si le test est significatif et si Xi>0, le spot est marqué « positif et significatif ». Les cas où le spot n'est pas « positif et significatif » sont traités ultérieurement (cf Surrogates). Le seuil de p-value utilisé pour le test est 0.01.

- **Well above SD** : Cette étape signale les spots déjà notés comme « positifs et significatifs » pour lesquels $X_i > w \times SDB_i^*$, où Xi est le signal après spatial detrend et SDBi* est l'erreur additive du modèle universel divisée par 2.6. Ici le paramètre utilisé est w = 13.

Surrogates

Cette option calcule, pour les mesures d'intensité des spots qui ne sont pas « positifs et significatifs » ou pour lesquels $X_i < SDB_i^*$, une mesure de la dispersion du bruit de fond qui va remplacer leur valeur : $SDB_i^* \times p-value$ (p-value du test de Student utilisé pour IsPosAndSignif).

Multiplicative detrend

Cet algorithme permet de corriger un effet spatial multiplicatif sur chacun des canaux. Ce type d'effet peut provenir de temps de réaction différents entre le centre et la périphérie de la lame et produit un effet « dôme ». Le principe est de diviser les valeurs d'intensité par le gradient de cet effet. On calcule pour cela une surface de signal sur la lame à partir des valeurs d'intensités des features possédant des réplicats (après normalisation par la moyenne pour chaque population de réplicats et exclusion des features ayant des valeurs d'intensité trop faibles). La surface est approchée localement par des polynômes d'ordre 2. La nouvelle valeur du signal, est la valeur du signal divisée par une estimation du gradient au point considéré.

B.1.6 Correction des biais liés aux fluorochromes

Cette étape doit permettre de corriger le biais lié à l'utilisation de deux fluorochromes différents.

Choix d'un ensemble de spots « de référence »

Ces spots serviront à calculer le facteur de normalisation mais tous les spots de la lame seront normalisés. Les spots utilisés sont les features non marqués comme non uniformes ou population outliers, non saturés (ici, pas plus de 50% des pixels saturés), non contrôles, positifs et significatifs qui vérifient : $CS = \frac{|\rho_R - \rho_G|}{n} \leq \tau$, où ρ_R et ρ_G sont les rangs de ces features pour chacun des deux canaux et n est le nombre de features vérifiant les premières conditions. Cette condition permet de n'utiliser pour la normalisation que les spots vérifiant une « tendance centrale ». La valeur de τ est par défaut de 0.05. Cette méthode permet de baser la normalisation sur un sous-ensemble de spots pour lesquels les rangs des intensités dans les deux conditions (pour un même spot) sont « proches » (ce sont donc les spots ayant le plus de chances de vérifier l'hypothèse nulle de non-régulation entre les deux conditions biologiques).

Facteur de normalisation

La méthode dite « linear and lowess » est utilisée pour normalisée les signaux. Les deux transformations suivantes sont appliquées :

- **Linear** : Estimation du biais des fluorochromes par régression linéaire (effectuée sur chaque canal (indépendamment)). Le signal est multiplié par une constante de manière à ce que la moyenne géométrique soit égale à 1000.
- **Lowess** : Estimation du biais des fluorochromes par régression locale (lowess : voir définition dans la partie spatial detrend). (régression locale pour trouver la tendance centrale des données, puis ajustement des données pour les centrer en 0).

Surrogates

C'est à ce niveau que l'on remplace les valeurs des surrogates (ie après avoir choisi les spots de référence pour la normalisation). Les surrogates ne sont pas normalisés.

Calcul du signal normalisé

Les nouvelles valeurs de signal normalisé sont calculées. Il est à noter que l'étape de normalisation modifie complètement l'échelle des intensités.

B.1.7 Calculs des ratios

Cette étape permet le calcul des log-ratios et des p-values associées. Les log-ratios résument l'information pour un spot et les p-values donnent une information sur la significativité de la différence d'expression entre les deux échantillons co-hybridés sur la puce.

Surrogates

Dans les cas particuliers suivants, le LogRatio = 0 et PValue = 1 :

- les 2 canaux sont des surrogates
- le canal Cy5 (rouge) est un surrogate et $\frac{I_r}{I_g} > 1$
- le canal Cy3 (vert) est un surrogate et $\frac{I_r}{I_g} < 1$

Calcul du log ratio, de l'erreur sur le log ratio et de la p-value

Dans les autres cas, le log ratio est calculé comme le logarithme en base 10 de l'intensité Cy5 sur l'intensité Cy3.

L'erreur sur le log ratio et la p-value sont ensuite calculées à partir du modèle d'erreur choisi. Pour le « Universal error model », les variances sur chaque intensités sont obtenues à partir du modèle. Ensuite, une déviation du log ratio par rapport à 0 est calculée : $\text{Xdev} = \frac{I_r - I_g}{\sqrt{\sigma_r^2 + \sigma_g^2}}$ (correspond à la statistique d'un t-test), puis une p-value est obtenue. Enfin, une erreur sur le log ratio est calculée par $\text{LogRatioError} = \frac{\text{LogRatio}}{\text{Xdev}}$.

Pour le « propagated error model », l'erreur sur le log ratio est calculée de la manière suivante à partir des variances et covariance calculées sur les pixels d'un spot pour les différents canaux : $\text{LogRatioError}^2 = \frac{1}{(\ln 10)^2}\left[\frac{\sigma_r^2}{I_r^2} + \frac{\sigma_g^2}{I_g^2} - 2\frac{\sigma_{rg}^2}{I_r \times I_g}\right]$. La déviation Xdev est déduite par la même formule que précédemment, puis la p-value est calculée.

B.1.8 Options du contrôle qualité

Des valeurs permettant d'évaluer la qualité des données de puces à ADN sont également fournies par Feature Extraction. Les options suivantes concernent ces contrôles qualité. Des spike-in sont utilisés (petits ARN d'une autre espèce présents avec un gradient de concentration et incorporés dans l'échantillon au moment du marquage). Le nombre minimal de réplicats utilisés pour les différentes statistiques du contrôle qualité est fixé à 5 par défaut et la p-value définissant les gènes considérés comme régulés pour le contrôle qualité est fixée à 0.01 par défaut

B.1.9 Génération des résultats

Les résultats sont générés dans un unique fichier texte et les images de format JPEG sont compressées d'un facteur 4 par défaut.

B.2 Modèle d'erreur de Rosetta Resolver

Le modèle d'erreur développé par Rosetta a inspiré celui utilisé par Feature Extraction sans être totalement équivalent. En pratique, il n'est pas utilisé pour calculer l'erreur sur les features mais sert uniquement à combiner des lames et à passer des features aux reporters, des reporters aux séquences et des séquences aux gènes. Seule cette dernière partie du modèle sera donc développée [92].

La méthode « Combining » permet de combiner des lames correspondant aux mêmes conditions de traitement pour rendre les résultats plus robustes. Soit n lames correspondant à un traitement.

Calcul du log ratio

Le nouveau log ratio est calculé par une moyenne pondérée par l'erreur de sorte que les lames ayant une erreur plus importante influent moins sur le résultat final : $\text{LogRatio} = \frac{\sum_{i=1}^{n} w_i \text{LogRatio}_i}{\sum_{i=1}^{n} w_i}$, où $w_i = \frac{1}{\sigma^2_{\text{LogRatio}_i}}$, où $\sigma^2_{\text{LogRatio}_i}$ est l'erreur sur le log ratio de la lame i.

Calcul de l'erreur sur le log ratio

Deux estimations de l'erreur sont possibles :

- **Propagated error** : erreur de population dérivée des erreurs individuelles de chaque lame σ_i^2 à partir de la formule suivante : $\sigma_p^2 = \frac{1}{\sum_{i=1}^{n} w_i} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$.
- **Scattered error** : variance empirique calculée à partir des observations individuelles sur chaque lame LogRatio_i : $\sigma_s^2 = \frac{1}{(n-1)\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i (\text{LogRatio}_i - \text{LogRatio})^2$.

La « propagated error » dépend uniquement de l'estimation de l'erreur par le modèle d'erreur. Elle n'est pas influencée par n mais peut être biaisée. La « scattered error » est une estimation non biaisée mais elle est très instable quand n est petit. De plus, la « propagated error » ne contient des informations que sur la variabilité technique due aux mesures, alors que la « scattered error » contient des informations concernant la variabilité technique et la variabilité biologique si les animaux considérés sont différents. Les deux erreurs sont donc combinées pour obtenir une estimation plus fiable : $\sigma_{lratio} = \frac{\sigma_p + (n-1)\sigma_s}{n}$, où σ_p est la « propagated error » et σ_s est la « scattered error ».

La méthode « Squeezing » permet de passer des features aux reporters, des reporters aux séquences et des séquences aux gènes. Le principe est globalement le même que pour le « Combining ».

Annexe C :

Résultats sur les données 22k

Les lames de souris Agilent 22k comportent environ 22000 sondes permettant de mesurer l'expression des gènes de la souris pour un échantillon biologique. Elles sont plus anciennes et moins complètes que les lames 4-44k.

C.1 Etude de variabilité technique des lames 22k

C.1.1 Design expérimental

Le design expérimental de cette étude de variabilité technique est le même que celui utilisé pour les lames 4-44k. Seuls les échantillons d'ARN diffèrent. L'étude sur les lames 22k est plus ancienne et ne disposait donc pas des échantillons des expériences PPAR. L'ARN utilisé provient de 2 échantillons de la lignée cellulaire KB/S23-500 traitées par un composé Servier anticancéreux (lignée de carcinome epidermoïde humain rendue résistante à ce même composé Servier). Six réplicats techniques ont été obtenus pour chaque échantillon biologique : trois jours de marquages et deux jours d'hybridation (Figure C.1).

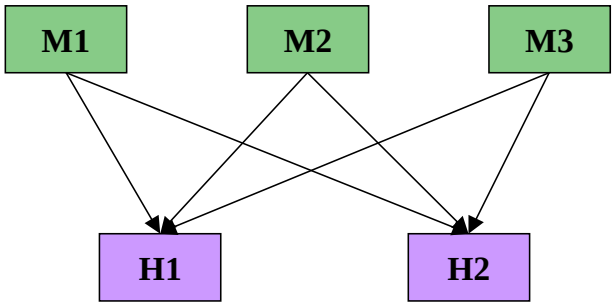

FIG. C.1 : Protocole de l'étude de variabilité technique
M = marquage, H = hybridation

Une procédure de « dye-swap » a été utilisée. Seules les séquences ayant une p-value inférieure ou égale à 0.05 pour au moins un réplicat technique ont été sélectionnées. Le bruit est défini de la même manière que dans le Chapitre 3 (cf 3.2.2) : le bruit sur le log ratio est la différence entre le log ratio de l'observation considérée et la moyenne des log ratios sur tous les réplicats techniques. Les bruits sur les intensités Cy3 et Cy5 sont définis de manière similaire.

C.1.2 Bruit sur le log ratio

La distribution du bruit sur le log ratio a la même allure que pour l'étude 4-44k et n'est pas normale (Figure C.2). Ceci est confirmé par un test de Shapiro-Wilk pour lequel des p-values significatives inférieures à 0.01 ont été obtenues sur différents échantillons de taille 50. L'hypothèse de normalité est donc rejetée. La dépendance entre bruit sur le log ratio et valeur du log ratio a également été étudiée. Il ne semble pas y avoir de relation exploitable entre ces deux valeurs (Figure C.3). Comme pour l'étude 4-44k, seuls quelques points montrent une relation linéaire entre bruit sur le log ratio et log ratio.

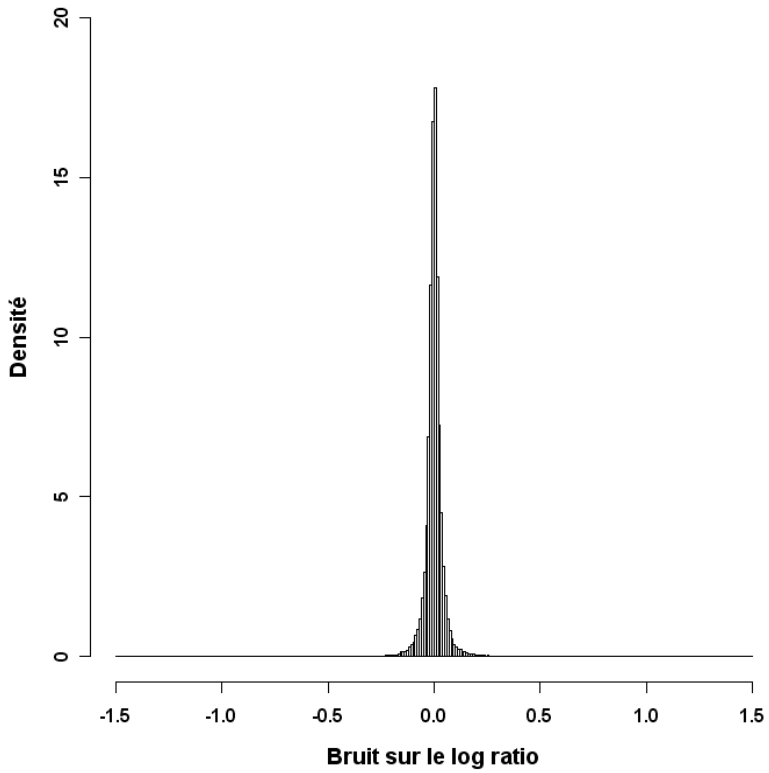

FIG. C.2 : Histogramme du bruit sur le log ratio
Les barres sont de largeur 0.01. Les deux échantillons biologiques sont regroupés.

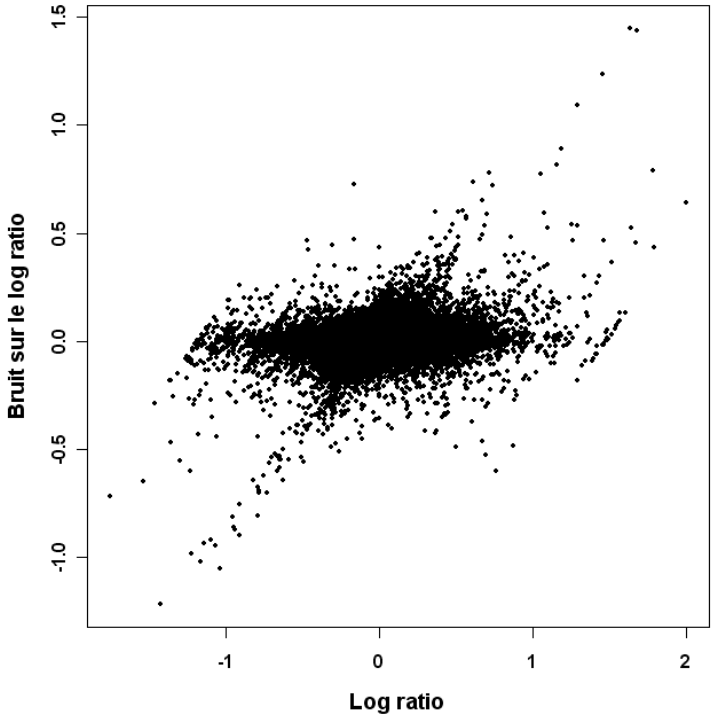

FIG. C.3 : Bruit sur le log ratio en fonction du log ratio
Les deux échantillons biologiques sont regroupés.

C.1.3 Lien avec les intensités

Le bruit sur les intensités a également été étudié pour les lames 22k. Comme pour le bruit sur le log ratio, on observe sur la Figure C.4.a que le bruit sur l'intensité ne suit pas une loi normale (p-values inférieures à 0.01 au test de Shapiro-Wilk sur des échantillons de taille 50). Le deuxième histogramme du bruit sur les intensités a été représenté pour des valeurs comprises entre –500 et 500, représentant environ 73% des valeurs initiales, afin de mieux discerner la forme de la distribution (Figure C.4.b).

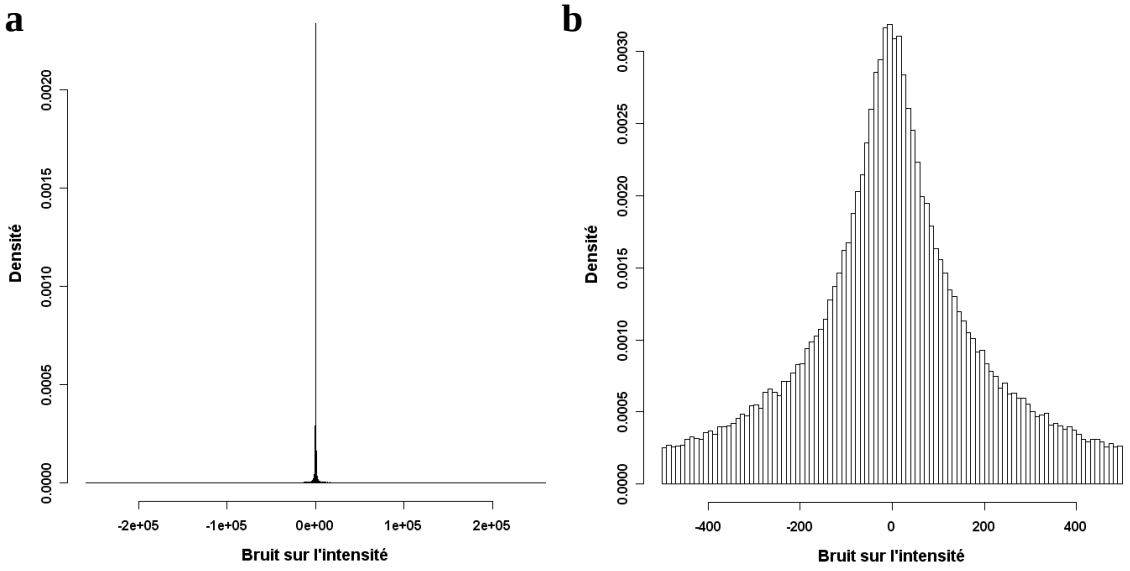

FIG. C.4 : Histogrammes du bruit sur l'intensité
(a) : Histogramme avec toutes les valeurs d'intensité. (b) : Histogramme avec les valeurs d'intensités comprises entre –500 et 500. Pour les deux histogrammes, les barres sont de largeur 10. Les deux échantillons biologiques et les deux intensités (Cy3 et Cy5) sont regroupés.

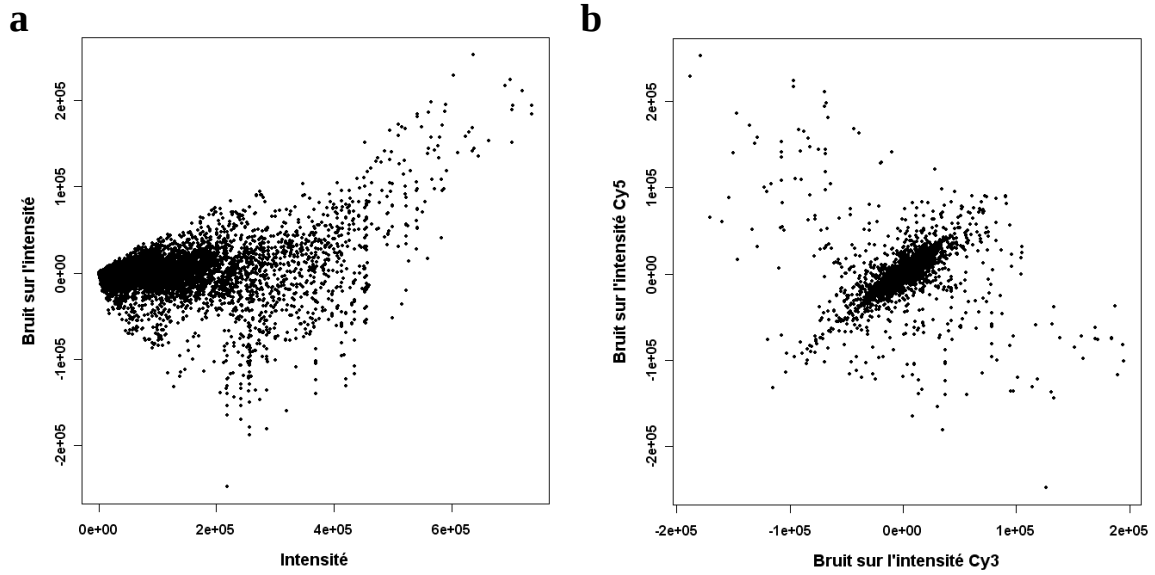

FIG. C.5 : Propriétés du bruit sur l'intensité
(a) : Bruit sur l'intensité en fonction de l'intensité. Les deux échantillons biologiques et les deux intensités (Cy3 et Cy5) sont regroupés. (b) : Bruit sur l'intensité Cy5n fonction du bruit sur l'intensité Cy3.

La relation entre bruit sur l'intensité et valeur de l'intensité est représentée sur la Figure C.5.a. L'écart type du bruit augmente avec l'intensité, comme pour les lames 4-44k. Enfin, il faut noter que les bruits sur les intensités Cy3 et Cy5 ne semblent pas complètement indépendants (Figure C.5.b). On peut noter qu'il y a plus de bruits décorrélés entre les deux intensités que pour les lames 4-44k. La relation entre bruit sur le log ratio et intensités a également été considérée (Figure C.6). On observe sur la Figure C.6.a que le bruit est beaucoup plus variable pour de faibles intensités. La Figure C.6.b, qui permet d'observer les intensités supérieures à 500 à une autre échelle, montre que l'écart type du bruit diminue dans un premier temps avec l'intensité, avant de réaugmenter fortement sur des hautes intensités.

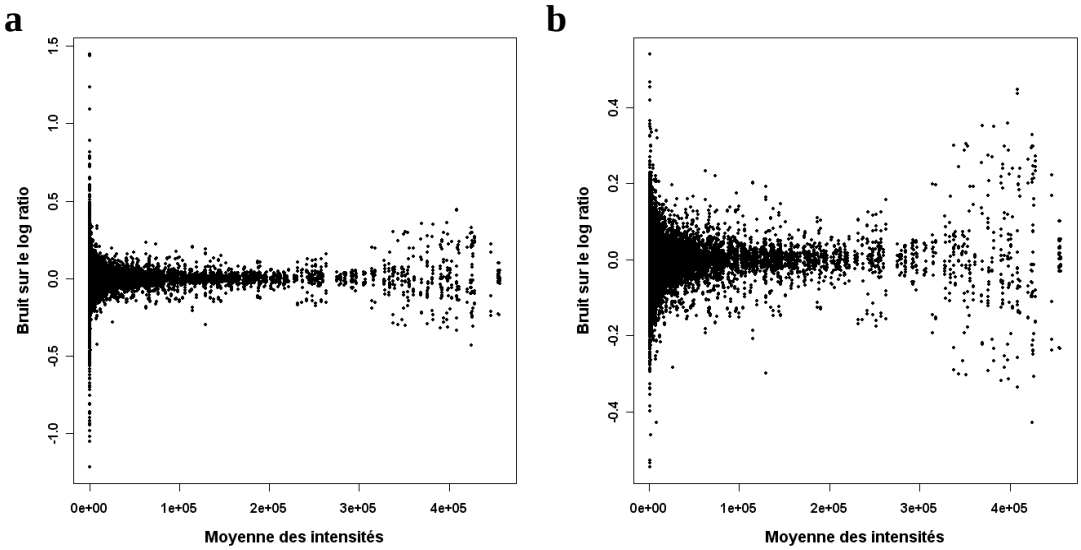

FIG. C.6 : Bruit sur le log ratio en fonction de la moyenne des intensités Cy3 et Cy5
(a) : Toutes les valeurs d'intensité. (b) : Valeurs d'intensité supérieures à 500. Pour les deux graphes, les valeurs d'intensité considérées ont également été moyennées sur les réplicats techniques et les deux échantillons biologiques sont regroupés.

Globalement, les résultats obtenus sont similaires à ceux de l'étude de variabilité technique sur lames 4-44k. La différence majeure est la diminution du bruit aux hautes intensités pour les lames 4-44k par rapport aux lames 22k.

C.2 Paramètres des méthodes

Les trois méthodes de sélection de variables présélectionnées ont également été paramétrées et testées sur les lames de souris 22k. De même que pour les lames 4-44k, les méthodes de T-test et de Nearest Shrunken Centroids ne nécessitent pas de paramétrage particulier puisqu'elles sont uniquement utilisées pour classer toutes les séquences. Les paramètres

déterminés pour la méthode de Support Vector Machines – Recursive Feature Elimination (SVM-RFE) sont présentés ici. D'autre part, une étude de la méthode des K plus proches voisins couplée à un algorithme génétique (KNN-GA) a également été réalisée, menant à l'abandon de cette méthode pour des raisons de non-reproductibilité et de temps de calcul. Tous ces tests ont été réalisés sur deux jeux de données 22k comparant les effets du tesaglitazar (agoniste mixte PPAR) au composé Servier SCOMP dans le foie et le tissu adipeux inguinal de souris ob/ob : TS_Foie et TS_TAI. Pour l'étude 22k, un seuil de p-value de 0.05 a été utilisé pour définir les séquences statistiquement significativement régulées.

C.2.1 Support Vector Machines – Recursive Feature Elimination

On rappelle que les paramètres à déterminer dans le cadre d'une SVM-RFE sont : le noyau du SVM, la vitesse speed (nombre de variables à partir duquel on ne retire qu'une seule variable à la fois) et le paramètre de soft-margin C.

3.3.3.1 Choix du noyau

Trois types de noyaux ont été testés : linéaire, polynomial (de degrés 2, 3, 4 et 5) et gaussien (d'écart type σ valant 0.1, 0.5, 1, 2 et 10). Les autres paramètres étaient fixés aux valeurs suivantes : $C = +\infty$, speed = 1000, $feat \in \{100, 500\}$. La qualité des résultats a été évaluée par l'erreur obtenue en Leave-One-Out-Cross-Validation (LOOCV) et est présentée dans le Tableau C.1. Le noyau linéaire obtient globalement les meilleurs résultats avec toutes les erreurs par LOOCV nulles. En revanche, quel que soit le degré utilisé, le noyau polynomial n'est pas très performant. Le noyau gaussien obtient une erreur faible uniquement pour un écart type à 10 et reste moins bon que le noyau linéaire. Le noyau linéaire est donc le plus performant sur ce type de données, comme pour les lames 4-44k.

	feat = 100		**feat = 500**	
Noyau	**TS_Foie**	**TS_TAI**	**TS_Foie**	**TS_TAI**
NL	0	0	0	0
NP degré 2	0.417	0.417	0.5	0.417
NP degré 3	0.167	0.333	0.083	0.083
NP degré 4	0.333	0.25	0.583	0.5
NP degré 5	0.167	0.083	0.25	0
NG s 0.1	1	1	1	1
NG s 0.5	1	1	1	1
NG s 1	1	1	1	1
NG s 2	0.417	0.417	1	1
NG s 10	0	0	0.083	0.083

TAB. C.1 : Erreurs par LOOCV obtenues avec différents types de noyaux
NL : noyau linéaire, NP : noyau polynomial, NG : noyau gaussien

3.3.3.2 Choix de C, paramètre de soft-margin

Différentes valeurs du paramètre de soft-margin C ont été testées afin d'étudier son impact sur chaque type de noyau : $C \in \{+\infty, 100, 10, 1, 0.1\}$. Un noyau linéaire, un noyau polynomial de degré 2 et un noyau gaussien d'écart type 1 ont été utilisés. Le paramètre feat a pris les valeurs 100 et 500, speed était fixé à 1000 et les performances ont été évaluées avec l'erreur par LOOCV. Quel que soit le type de noyau, les résultats ne sont pas modifiés par la valeur de C. Il semble donc pertinent de choisir la valeur de C la plus restrictive sur la qualité de la classification, c'est-à-dire $C = +\infty$. Les listes de séquences obtenues avec les différentes valeurs de C ont donc été comparées à celle obtenue avec $C = +\infty$ afin de vérifier que les listes n'étaient pas trop différentes (Tableau C.2). On observe qu'il y très peu de différences entre $C = +\infty$ et les autres valeurs pour $feat = 500$. La valeur de $C = +\infty$ a donc été conservée, comme pour l'étude sur les lames 4-44k.

	feat = 100		**feat = 500**	
C	**TS_Foie**	**TS_TAI**	**TS_Foie**	**TS_TAI**
100	93	88	99.4	98.8
10	91	89	100	99.8
1	100	100	100	100
0.1	91	87	100	99.6

TAB. C.2 : Pourcentage de séquences communes entre les listes obtenues avec les différentes valeurs de C et celle obtenue avec $= +\infty$

3.3.3.3 Choix de speed, paramètre de vitesse

En dernier lieu, le paramètre speed a été considéré. Avec un noyau linéaire, $C = +\infty$ et feat valant 100 et 500, quatre valeurs de speed ont été testées. Pour chaque jeu de données test, ces valeurs ont été choisies de manière à diviser le nombre de séquences par deux, 0 fois, 1 fois, 2 fois ou 3 fois. Les erreurs par LOOCV obtenues sont toutes nulles peu importe la valeur du paramètre. Les listes de séquences obtenues avec les différentes valeurs de speed ont également été comparées à celle obtenue avec la valeur de speed correspondant à zéro division, c'est-à-dire à la meilleure précision (Tableau C.3). On observe que seule la valeur de speed correspondant à diviser 1 fois les séquences par deux ne modifie pas les résultats. C'est donc cette valeur qui a été choisie.

	feat = 100		feat = 500	
speed	**TS_Foie**	**TS_TAI**	**TS_Foie**	**TS_TAI**
division 1 fois	100	100	100	100
division 2 fois	91	88	95.2	89.8
division 3 fois	79	83	92.6	89

TAB. C.3 : Pourcentage de séquences communes entre les listes obtenues avec les différentes valeurs de speed et celle obtenue avec la valeur de speed correspondant à zéro division par deux du nombre de séquences

C.2.2 K plus proches voisins couplés à un algorithme génétique

Une méthode de sélection de variables couplant K plus proches voisins et algorithmes génétiques (KNN-GA) a également été sélectionnée dans la littérature avant d'être rejetée. Elle a été paramétrée sur les mêmes jeux de données que la SVM-RFE (TS_Foie et TS_TAI). Cette méthode consiste à trouver un groupe de séquences optimal au sens de la classification par la méthode des KNN. Un algorithme génétique est utilisé pour la phase d'optimisation. Pour des raisons de clarté, seuls les principaux résultats sont présentés ici : paramètres de la KNN-GA qui ont été utilisés et résultats justifiant le rejet de cette méthode. Les packages Spider [63] et GATBX [93] de Matlab ont respectivement été utilisés pour la méthode de KNN et l'algorithme génétique.

C.2.2.1 Paramètres de la KNN-GA

Les méthodes de K plus proches voisins et d'algorithmes génétiques sont présentées dans le Chapitre 2 de ce manuscrit (cf 2.2.1.4). Pour rappel, un algorithme génétique est une méthode d'optimisation qui fait évoluer une population de solutions en suivant des règles proches de la génétique. Différents paramètres ont besoin d'être déterminés : définition d'une population de solutions, fonction score à optimiser, méthode de croisement entre deux individus, méthode de mutation d'un individu et enfin critère d'arrêt de l'algorithme.

Dans notre étude, une population est un ensemble de 40 individus qui sont définis de la manière suivante. Chaque individu correspond à un groupe de séquences et est représenté par une chaîne de 0 et de 1 de longueur égale au nombre total de variables. Cette chaîne prend la valeur 1 pour une séquence appartenant au groupe et 0 sinon. Différentes tailles de groupes de séquences ont été testées : 100, 150 et 200. Les individus ont été évalués par leur qualité de classification avec une méthode des K plus proches voisins (K = 5, distance euclidienne).

Pour cela, une fonction score F(ind) a été définie pour chaque individu ind : $F(ind) = \delta(ind) + C\left|d_0^2(ind) - d_1^2(ind)\right|$, où $\delta(ind)$ vaut 1 si les K plus proches voisins de ind ont la même classe que lui et 0 sinon. d_j est la distance de l'individu ind à la classe j (distance minimale) et C est un critère de séparabilité des classes qui vaut 10^{-4} [42]. C'est cette fonction score qui a été optimisée par l'algorithme génétique.

36 individus sont tirés avec une probabilité dépendant de leur fonction de score (tirage avec remise). Ils sont ensuite croisés deux par deux, puis mutés avec une probabilité 10^{-4} de manière à générer 36 nouveaux individus. Un croisement correspond à échanger des séquences entre deux individus et une mutation consiste à modifier des séquences au sein d'un individu. La nouvelle population est constituée de ces 36 nouveaux individus et des 4 meilleurs individus de l'ancienne population. L'algorithme s'arrête quand la fonction score n'a pas été améliorée de plus de 10^{-4} pendant MaxGen itérations ou quand il atteint les 10000 itérations. MaxGen dépend du nombre de séquences L dans les groupes : $MaxGen = 100 + 2 \times L$. Quand l'algorithme est terminé, l'individu ayant la meilleure valeur de fonction score est conservé.

Enfin, les algorithmes génétiques sont caractérisés par leur caractère aléatoire. Différents lancements peuvent mener, pour une même problématique, à des résultats complètement différents. Ceci est essentiellement lié à la multiplicité de choix de la population initiale de solutions. Afin de tenter de compenser ce problème, une étude de fréquence a été réalisée. La liste de séquences conservée pour la méthode de KNN-GA est constituée des séquences ayant été sélectionnées dans au moins 33% de N lancements de l'algorithme. N a pris les valeurs 12, 20, 30 et 60.

C.2.2.2 Inconvénients de la KNN-GA

Cette méthode de KNN-GA présente néanmoins plusieurs inconvénients qui ont conduit à son élimination de l'étude. D'une part, un certain nombre de lancements sont nécessaires afin d'espérer obtenir des résultats reproductibles (Tableau C.4). On observe en effet qu'une reproductibilité acceptable (supérieure à 90%) n'est atteinte pour le jeu de données TS_TAI qu'en considérant au moins N = 30 lancements pour 200 séquences et au moins N = 60 pour 150 séquences. En considérant qu'un lancement dure 21 minutes sur TS_Foie pour 200 séquences, la méthode de KNN-GA est très coûteuse (PC de 2.4GHz et 2Go de RAM). Pour comparaison, un lancement de SVM-RFE dure environ 6 minutes mais ne nécessite pas d'étude de fréquence et un lancement du T-test ou de la méthode NSC prend moins d'une minute.

Nombre de lancements	Nombre de séquences	TS_Foie	TS_TAI
12	100	81.16	73.46
	150	88.05	82.02
	200	91.6	87.28
20	100	82.78	78.07
	150	90.34	85.6
	200	93.43	89.57
30	100	85.66	81.85
	150	92.56	87.78
	200	93.97	91.29
60	100	89.92	86.78
	150	97.89	91.96
	200	94.87	93.76

TAB. C.4 : Reproductibilité des listes de séquences obtenues avec la KNN-GA pour différents nombres de lancements et pour différents nombres de séquences
La reproductibilité est le pourcentage moyen de gènes communs entre plusieurs runs.

D'autre part, le nombre de séquences doit être fixé à l'avance. Pour pouvoir déterminer la bonne longueur de liste de séquences, en calculant par exemple la robustesse à des longueurs différentes avec MetRob, il faudrait donc répéter la KNN-GA un nombre M de fois. Le temps de calcul serait alors encore considérablement augmenté ($25\ \text{min} \times 30\ \text{lancements} \times \text{M}$). La méthode de KNN-GA a donc été abandonnée avant même d'être testée sur les données obtenues avec des lames 4-44k.

Annexe D :

Précisions d'ordre mathématique

D.1 Test de Shapiro-Wilk

Le test de Shapiro-Wilk est un test statistique qui permet de vérifier si un échantillon de données est issu d'une distribution normale [94]. Soit $x = (x_1,\ldots,x_n)$ l'échantillon considéré. Notons H_0 et H_1 les deux hypothèses suivantes :

- H_0 : x est issu d'une distribution normale.
- H_1 : x n'est pas issu d'une distribution normale.

Soit W la statistique du test, $W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, où :

- $x_{(i)}$ est la i-ème valeur de x par ordre croissant.
- $\bar{x}$ est la moyenne empirique de x.
- $(a_1,\ldots,a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$, avec $(m_1,\ldots,m_n)$ tirées de manière indépendante et identiquement distribuée à partir d'une distribution normale puis classées par ordre croissant et V leur matrice de covariance.

W est interprété comme un coefficient de corrélation quadratique entre la distribution observée et une distribution normale. La p-value obtenue avec ce test représente la probabilité de rejeter à tort l'hypothèse H_0. Le test de Shapiro-Wilk devient très sensible pour des valeurs de n élevées et a donc été appliqué à des échantillons de taille 50. La fonction shapiro.test du logiciel R a été utilisée.

D.2 Moments d'une loi normale au carrée signée

Soit V, variable aléatoire telle que $V = \text{signe}(U) \times U^2$, où U suit une loi normale centrée en zéro et d'écart type σ. On cherche à calculer l'espérance et l'écart type de V. Soit f la densité de probabilité de U, alors $f(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u}{\sigma}\right)^2}$. On a $V = H(U) = \text{signe}(U) \times U^2$ avec H croissante et bijective, donc si g est la densité de probabilité de V, on obtient $g(v) = \frac{f\left[H^{-1}(v)\right]}{H'[H^{-1}(v)]}$. Comme $H^{-1}(v) = \sqrt{|v|}$ et $H'(v) = 2|v|$, alors $g(v) = \frac{1}{\sigma\sqrt{2\pi} \times 2|v|} e^{-\frac{1}{2\sigma^2}|v|}$.

Les fonctions $v \to vg(v)$ et $v \to v^2 g(v)$ sont intégrables sur IR. Ainsi, l'espérance de V est donnée par :

$$\begin{aligned} E(V) &= \int_{-\infty}^{+\infty} vg(v)dv \\ &= \int_0^{+\infty} \frac{\sqrt{v}}{2\sqrt{2\pi}\sigma} e^{-\frac{v}{2\sigma^2}} dv + \int_{-\infty}^{0} -\frac{\sqrt{-v}}{2\sqrt{2\pi}\sigma} e^{\frac{v}{2\sigma^2}} dv \\ &= \int_0^{+\infty} \frac{\sqrt{v}}{2\sqrt{2\pi}\sigma} e^{-\frac{v}{2\sigma^2}} dv - \int_0^{+\infty} \frac{\sqrt{v}}{2\sqrt{2\pi}\sigma} e^{-\frac{v}{2\sigma^2}} dv \\ &= 0 \end{aligned}$$

D'autre part, la variance de V vérifie :

$$\begin{aligned} \text{Var}(V) &= \int_{-\infty}^{+\infty} v^2 g(v)dv \\ &= \int_0^{+\infty} \frac{v\sqrt{v}}{2\sqrt{2\pi}\sigma} e^{-\frac{v}{2\sigma^2}} dv + \int_{-\infty}^{0} -\frac{v\sqrt{v}}{2\sqrt{2\pi}\sigma} e^{-\frac{v}{2\sigma^2}} dv \\ &= 2\int_0^{+\infty} \frac{v\sqrt{v}}{2\sqrt{2\pi}\sigma} e^{-\frac{v}{2\sigma^2}} dv \end{aligned}$$

Enfin, en appliquant le changement de variable $z = \sqrt{v}$, on obtient :

$$\begin{aligned}
\mathrm{Var}(V) &= 2\int_0^{+\infty} \frac{z^4}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}} dz \\
&= \left[-\frac{\sigma z^3}{\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \left(-\frac{3\sigma\sigma^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} dz \right) \\
&= \int_{-\infty}^{+\infty} \frac{3\sigma^4 z^4}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= 3\sigma^4
\end{aligned}$$

Ainsi, si on note Σ l'écart type de V, on a la relation suivante : $\Sigma^2 = 3\sigma^4$.

D.3 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test statistique qui peut être utilisé pour vérifier si deux échantillons sont issus d'une même distribution [95]. Soient $x = (x_1,\ldots,x_n)$ et $y = (y_1,\ldots,y_m)$ deux échantillons. Notons H_0 et H_1 les deux hypothèses suivantes :

- H_0 : les deux échantillons sont issus d'une même distribution.
- H_1 : les deux échantillons ne sont pas issus d'une même distribution.

Définissons $F_{xn}(u)$ la fonction de répartition empirique de x : $F_{xn}(u) = \frac{1}{n}\sum_{i=1}^{n} I_{x_i \le u}$, avec $I_{x_i \le u} = \begin{cases} 1 \text{ si } x_i \le u \\ 0 \text{ sinon} \end{cases}$. $F_{ym}(u)$ est défini de manière similaire. Soit $D_{n,m} = \underset{u \in IR}{\mathrm{Sup}}\left(\left|F_{xn}(u) - F_{ym}(u)\right|\right)$, la statistique de Kolmogorov-Smirnov. Plus $D_{n,m}$ est grand, plus les distributions des deux échantillons sont éloignées. L'hypothèse H_0 est testée en comparant $\sqrt{\frac{nm}{n+m}} D_{n,m}$ à une valeur critique K_α dépendant du risque α accepté (probabilité de rejeter à tort H_0). Si $\sqrt{\frac{nm}{n+m}} D_{n,m} > K_\alpha$, l'hypothèse H_0 est rejeté avec une confiance $1-\alpha$.

La fonction ks.test du logiciel R a été utilisée pour le test de Kolmogorov-Smirnov. Pour des échantillons de grande taille (>10000) ou dans les cas d'égalité entre deux valeurs, la p-value fournie est une approximation.

D.4 Test exact de Fisher

Le test exact de Fisher est un test statistique qui permet entre autres de tester si les probabilités d'appartenance à une classe sont identiques entre deux populations [96]. Notons H_0 et H_1 les deux hypothèses suivantes :

- H_0 : les probabilités d'appartenance à une classe sont identiques entre les deux populations.
- H_1 : les probabilités d'appartenance à une classe ne sont pas identiques entre les deux populations.

	Population 1	**Population 2**	
Classe 1	a	b	a + b
Classe 2	c	d	c + d
	a + c	b + d	n

TAB. D.1 : Table de contingence de deux populations divisées en deux classes
a (respectivement c) est le nombre d'individus de la population 1 appartenant à la classe 1 (respectivement 2). b (respectivement d) est le nombre d'individus de la population 2 appartenant à la classe 1 (respectivement 2). n est le nombre total d'individus.

Sous l'hypothèse nulle, la probabilité d'obtenir une telle table est donnée par la distribution hypergéométrique : $p = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_{n}^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$ [97]. La p-value obtenue avec ce test représente la probabilité de rejeter à tort l'hypothèse H_0. La fonction fisher.test du logiciel R a été utilisée.

Bibliographie

[1] *Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia*, World Health Organization , 2006.

[2] "WHO | Diabetes : http://www.who.int/diabetes/en/."

[3] Vivant Editions, *Dossiers d'Actualités: Diabète*, Pôle santé parisien, 2006.

[4] J.M. Berg, J.L. Tymoczko, et L. Stryer, *Biochemistry, Fifth Edition: International Version*, W. H. Freeman, 2002.

[5] I. CUSIN et F. ROHNER-JEANRENAUD, "Boucle régulatrice entre le neuropeptide Y et la leptine et son altération chez le rongeur obèse: Obésité: fondements expérimentaux de nouvelles thérapeutiques," *MS. Médecine sciences*, vol. 14, 1998, pp. 907-913.

[6] S.R. Votey et A.L. Peters, "Diabetes Mellitus, Type1 - A Review," *Emedecine*, 2009.

[7] T. Guo et M. Hebrok, "Stem cells to pancreatic beta-cells: new sources for diabetes cell therapy," *Endocrine Reviews*, vol. 30, Mai. 2009, pp. 214-227.

[8] Z. Yang, M. Chen, J.D. Carter, C.S. Nunemaker, J.C. Garmey, S.D. Kimble, et J.L. Nadler, "Combined treatment with lisofylline and exendin-4 reverses autoimmune diabetes," *Biochemical and Biophysical Research Communications*, vol. 344, Jun. 2006, pp. 1017-1022.

[9] S.R. Votey et A.L. Peters, "Diabetes Mellitus, Type2 - A Review," *Emedecine*, 2009.

[10] S.E. Inzucchi, "Oral antihyperglycemic therapy for type 2 diabetes: scientific review," *JAMA: The Journal of the American Medical Association*, vol. 287, Jan. 2002, pp. 360-372.

[11] H. Yki-Järvinen, "Thiazolidinediones," *The New England Journal of Medicine*, vol. 351, Sep. 2004, pp. 1106-18.

[12] G.M. Keating et K.F. Croom, "Fenofibrate: a review of its use in primary dyslipidaemia, the metabolic syndrome and type 2 diabetes mellitus," *Drugs*, vol. 67, 2007, pp. 121-53.

[13] D.M. Nathan, J.B. Buse, M.B. Davidson, E. Ferrannini, R.R. Holman, R. Sherwin, et B. Zinman, "Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes," *Diabetes Care*, vol. 32, Jan. 2009, pp. 193-203.

[14] W.T. Cefalu, "Animal models of type 2 diabetes: clinical presentation and pathophysiological relevance to the human condition," *ILAR Journal / National Research Council, Institute of Laboratory Animal Resources*, vol. 47, 2006, pp. 186-198.

[15] K. Srinivasan et P. Ramarao, "Animal models in type 2 diabetes research: an overview," *The Indian Journal of Medical Research*, vol. 125, Mar. 2007, pp. 451-72.

[16] E. Shafrir, E. Ziv, et R. Kalman, “Nutritionally induced diabetes in desert rodents as models of type 2 diabetes: Acomys cahirinus (spiny mice) and Psammomys obesus (desert gerbil),” *ILAR Journal / National Research Council, Institute of Laboratory Animal Resources*, vol. 47, 2006, pp. 212-224.
[17] D. Chen et M. Wang, “Development and application of rodent models for type 2 diabetes,” *Diabetes, Obesity & Metabolism*, vol. 7, Jul. 2005, pp. 307-317.
[18] K. Kobayashi, T.M. Forte, S. Taniguchi, B.Y. Ishida, K. Oka, et L. Chan, “The db/db mouse, a model for diabetic dyslipidemia: molecular characterization and effects of Western diet feeding,” *Metabolism: Clinical and Experimental*, vol. 49, Jan. 2000, pp. 22-31.
[19] B. Desvergne et W. Wahli, “Peroxisome proliferator-activated receptors: nuclear control of metabolism,” *Endocrine Reviews*, vol. 20, Oct. 1999, pp. 649-88.
[20] P.S. Jones, R. Savory, P. Barratt, A.R. Bell, T.J. Gray, N.A. Jenkins, D.J. Gilbert, N.G. Copeland, et D.R. Bell, “Chromosomal localisation, inducibility, tissue-specific expression and strain differences in three murine peroxisome-proliferator-activated-receptor genes,” *European Journal of Biochemistry / FEBS*, vol. 233, Oct. 1995, pp. 219-26.
[21] D. Duran-Sandoval, A.C. Thomas, B. Bailleul, J.C. Fruchart, et B. Staels, “Pharmacologie des agonistes de PPARa et PPARg et des activateurs PPARa/g mixtes en développement clinique,” *M/S: médecine sciences*, vol. 19, 2003, pp. 819-825.
[22] A.V. Schwartz, “TZDs and Bone: A Review of the Recent Clinical Evidence,” *PPAR Research*, vol. 2008, 2008, p. 297893.
[23] S.E. Nissen, K. Wolski, et E.J. Topol, “Effect of muraglitazar on death and major adverse cardiovascular events in patients with type 2 diabetes mellitus,” *JAMA: The Journal of the American Medical Association*, vol. 294, Nov. 2005, pp. 2581-6.
[24] F. Zhang, B.E. Lavan, et F.M. Gregoire, “Selective Modulators of PPAR-gamma Activity: Molecular Aspects Related to Obesity and Side-Effects,” *PPAR Research*, vol. 2007, 2007, p. 32696.
[25] V.S. Gomase, S. Tagore, et K.V. Kale, “Microarray: an approach for current drug targets,” *Current Drug Metabolism*, vol. 9, Mar. 2008, pp. 221-31.
[26] A. Agilent Technologies, “Agilent Whole Mouse Genome Oligo Microarray Kit with SurePrint Technology,” 2004.
[27] Rosetta Inpharmatics, *Rosetta Resolver Analysis Guide*, 2006.
[28] Agilent Technologies, *Agilent Feature Extraction Software (v9.5) Reference Guide*, 2007.
[29] J. Quackenbush, “Microarray data normalization and transformation,” *Nature Genetics*, vol. 32 Suppl, Déc. 2002, pp. 496-501.
[30] “Ingenuity Pathway Analysis Software-Complete Pathways Database : http://www.ingenuity.com/.”
[31] M. Kolak, H. Yki-Järvinen, K. Kannisto, M. Tiikkainen, A. Hamsten, P. Eriksson, et R.M. Fisher, “Effects of chronic rosiglitazone therapy on gene expression in human adipose tissue in vivo in patients with type 2 diabetes,” *The Journal of Clinical Endocrinology and Metabolism*, vol. 92, Fév. 2007, pp. 720-724.

[32] A. Bar-Hen et T. Mary-Huard, “Cours d'Analyse de données,” *Master 2 de probabilités et statistiques*, Orsay: 2008.
[33] T. Hastie, R. Tibshirani, et J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag New York Inc., 2003.
[34] Y. Saeys, I. Inza, et P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics (Oxford, England)*, vol. 23, Oct. 2007, pp. 2507-17.
[35] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *IJCAI*, 1995, pp. 1145, 1137.
[36] C. Ambroise et G.J. McLachlan, “Selection bias in gene extraction on the basis of microarray gene-expression data,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, Mai. 2002, pp. 6562-6.
[37] T. Li, C. Zhang, et M. Ogihara, “A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression,” *Bioinformatics (Oxford, England)*, vol. 20, Oct. 2004, pp. 2429-37.
[38] R. Dudoit, J. Fridly, et T.P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, vol. 97, 2002, pp. 77--87.
[39] E.P. Xing, M.I. Jordan, et R.M. Karp, “Feature selection for high-dimensional genomic microarray data,” *IN PROCEEDINGS OF THE EIGHTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, 2001, pp. 601--608.
[40] L. Li, T.A. Darden, C.R. Weinberg, A.J. Levine, et L.G. Pedersen, “Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method,” *Combinatorial Chemistry & High Throughput Screening*, vol. 4, Déc. 2001, pp. 727-39.
[41] T. Jirapech-Umpai et S. Aitken, “Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes,” *BMC Bioinformatics*, vol. 6, 2005, p. 148.
[42] J.M. Deutsch, “Evolutionary algorithms for finding optimal gene sets in microarray prediction,” *Bioinformatics (Oxford, England)*, vol. 19, Jan. 2003, pp. 45-52.
[43] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, et S. Levy, “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis,” *Bioinformatics (Oxford, England)*, vol. 21, Mar. 2005, pp. 631-43.
[44] “Fichier:ArtificialNeuronModel francais.png - Wikipédia : http://fr.wikipedia.org/wiki/Fichier:ArtificialNeuronModel_francais.png.”
[45] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, et P.S. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature Medicine*, vol. 7, Jun. 2001, pp. 673-9.
[46] O. Bousquet, “Introduction aux “Support Vector Machines”,” 2001.
[47] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, et T.R. Golub, “Multiclass cancer diagnosis using tumor gene expression signatures,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, Déc. 2001, pp. 15149-54.

[48] F. Chu et L. Wang, “Applications of support vector machines to cancer classification with microarray data,” *International Journal of Neural Systems*, vol. 15, Déc. 2005, pp. 475-84.
[49] R. Tibshirani, T. Hastie, B. Narasimhan, et G. Chu, “Diagnosis of multiple cancer types by shrunken centroids of gene expression,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, Mai. 2002, pp. 6567-72.
[50] V.G. Tusher, R. Tibshirani, et G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, Avr. 2001, pp. 5116-5121.
[51] S. Zhang, “A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance,” *BMC Bioinformatics*, vol. 8, 2007, p. 230.
[52] R. Bijlani, Y. Cheng, D.A. Pearce, A.I. Brooks, et M. Ogihara, “Prediction of biologically significant components from microarray data: Independently Consistent Expression Discriminator (ICED),” *Bioinformatics (Oxford, England)*, vol. 19, Jan. 2003, pp. 62-70.
[53] X. Li, S. Rao, Y. Wang, et B. Gong, “Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling,” *Nucleic Acids Research*, vol. 32, 2004, pp. 2685-94.
[54] I. Guyon, J. Weston, S. Barnhill, et V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” *Machine Learning*, vol. 46, 2002, pp. 389–422.
[55] L.M. Fu et C.S. Fu-Liu, “Evaluation of gene importance in microarray data based upon probability of selection,” *BMC Bioinformatics*, vol. 6, 2005, p. 67.
[56] “Algorithmes génétiques : http://lsis.univ-tln.fr/~tollari/TER/AlgoGen1/node5.html.”
[57] S. Peng, Q. Xu, X.B. Ling, X. Peng, W. Du, et L. Chen, “Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines,” *FEBS Letters*, vol. 555, Déc. 2003, pp. 358-62.
[58] L. Li, W. Jiang, X. Li, K.L. Moser, Z. Guo, L. Du, Q. Wang, E.J. Topol, Q. Wang, et S. Rao, “A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset,” *Genomics*, vol. 85, Jan. 2005, pp. 16-23.
[59] C.F. Aliferis, I. Tsamardinos, et A. Statnikov, “HITON: a novel Markov Blanket algorithm for optimal variable selection,” *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 2003, pp. 21-5.
[60] L. Guo, E.K. Lobenhofer, C. Wang, R. Shippy, S.C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F.M. Goodsaid, P. Hurban, K.L. Phillips, J. Xu, X. Deng, Y.A. Sun, W. Tong, Y.P. Dragan, et L. Shi, “Rat toxicogenomic study reveals analytical consistency across microarray platforms,” *Nature Biotechnology*, vol. 24, Sep. 2006, pp. 1162-1169.
[61] A. Cotillard, S. Le Bouter, N. Guigal-Stéphan, S. Courtade-Gaïani, C. Dacquet, M. Lonchampt, A. Ktorza, F. Xavier, B. Lockhart, et J. Galizzi, “NSCRob, a novel approach based on Nearest Shrunken Centroids to select lists of genes with improved robustness: transcriptomic comparison of two Peroxisome Proliferator-Activated Receptor agonists in db/db mice,” *BMC Bioinformatics*, Submitted. .
[62] “PAM : Prediction Analysis for Microarrays : http://www-stat.stanford.edu/~tibs/PAM/.”

[63] “Spider : http://www.kyb.mpg.de/bs/people/spider/.”
[64] “RPy home page : http://rpy.sourceforge.net/.”
[65] “mlabwrap : http://mlabwrap.sourceforge.net/.”
[66] L.M. McShane, M.D. Radmacher, B. Freidlin, R. Yu, M. Li, et R. Simon, “Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data,” *Bioinformatics (Oxford, England)*, vol. 18, Nov. 2002, pp. 1462-9.
[67] A. Sayyed-Ahmad, K. Tuncay, et P.J. Ortoleva, “Transcriptional regulatory network refinement and quantification through kinetic modeling, gene expression microarray data and information theory,” *BMC Bioinformatics*, vol. 8, 2007, p. 20.
[68] “NCTR Center for Toxicoinformatics - MAQC Project : http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/.”
[69] “CRAN - Package qvalue : http://cran.r-project.org/web/packages/qvalue/index.html.”
[70] C.J. Albers, R.C. Jansen, J. Kok, O.P. Kuipers, et S.A. van Hijum, “SIMAGE: simulation of DNA-microarray gene expression data,” *BMC Bioinformatics*, vol. 7, 2006, p. 205.
[71] “Bioinformatics @ MolGen : SIMAGE : http://bioinformatics.biol.rug.nl/websoftware/simage/simage_start.php.”
[72] S. Scheid, C. Lottaz, X. Yang, et R. Spang, *Similarities of Ordered Gene Lists - User’s Guide to the Bioconductor Package OrderedList*, CompDiag, 2006.
[73] “Umetrics - SIMCA-P for Multivariate Data Analysis : http://www.umetrics.com/default.asp/pagename/software_simcap/c/3.”
[74] S.M. Watkins, P.R. Reifsnyder, H. Pan, J.B. German, et E.H. Leiter, “Lipid metabolome-wide effects of the PPARgamma agonist rosiglitazone,” *Journal of Lipid Research*, vol. 43, Nov. 2002, pp. 1809-1817.
[75] I. García-Ruiz, C. Rodríguez-Juan, T. Díaz-Sanjuán, M.A. Martínez, T. Muñoz-Yagüe, et J.A. Solís-Herruzo, “Effects of rosiglitazone on the liver histology and mitochondrial function in ob/ob mice,” *Hepatology (Baltimore, Md.)*, vol. 46, Aoû. 2007, pp. 414-423.
[76] A.M. Sharma et B. Staels, “Review: Peroxisome proliferator-activated receptor gamma and adipose tissue--understanding obesity-related changes in regulation of lipid and glucose metabolism,” *The Journal of Clinical Endocrinology and Metabolism*, vol. 92, Fév. 2007, pp. 386-395.
[77] T.M. Chan, K.M. Young, N.J. Hutson, F.T. Brumley, et J.H. Exton, “Hepatic metabolism of genetically diabetic (db/db) mice. I. Carbohydrate metabolism,” *The American Journal of Physiology*, vol. 229, Déc. 1975, pp. 1702-1712.
[78] N.H. Jeoung et R.A. Harris, “Pyruvate dehydrogenase kinase-4 deficiency lowers blood glucose and improves glucose tolerance in diet-induced obese mice,” *American Journal of Physiology. Endocrinology and Metabolism*, vol. 295, Jul. 2008, pp. E46-54.
[79] M. Loffler, M. Bilban, M. Reimers, W. Waldhäusl, et T.M. Stulnig, “Blood glucose-lowering nuclear receptor agonists only partially normalize hepatic gene expression in db/db mice,” *The Journal of Pharmacology and Experimental Therapeutics*, vol. 316, Fév. 2006, pp. 797-804.
[80] J.X. Rong, Y. Qiu, M.K. Hansen, L. Zhu, V. Zhang, M. Xie, Y. Okamoto, M.D. Mattie, H. Higashiyama, S. Asano, J.C. Strum, et T.E. Ryan, “Adipose mitochondrial biogenesis

is suppressed in db/db and high-fat diet-fed mice and improved by rosiglitazone," *Diabetes*, vol. 56, Jul. 2007, pp. 1751-1760.
[81] C.D. Toseland, S. Campbell, I. Francis, P.J. Bugelski, et N. Mehdi, "Comparison of adipose tissue changes following administration of rosiglitazone in the dog and rat," *Diabetes, Obesity & Metabolism*, vol. 3, Jun. 2001, pp. 163-170.
[82] S.P. Weisberg, D. McCann, M. Desai, M. Rosenbaum, R.L. Leibel, et A.W. Ferrante, "Obesity is associated with macrophage accumulation in adipose tissue," *The Journal of Clinical Investigation*, vol. 112, Déc. 2003, pp. 1796-1808.
[83] M. Yousef, M. Ketany, L. Manevitz, L.C. Showe, et M.K. Showe, "Classification and biomarker identification using gene network modules and support vector machines," *BMC Bioinformatics*, vol. 10, 2009, p. 337.
[84] L. Song, J. Bedo, K.M. Borgwardt, A. Gretton, et A. Smola, "Gene selection via the BAHSIC family of algorithms," *Bioinformatics (Oxford, England)*, vol. 23, Jul. 2007, pp. i490-498.
[85] I.B. Jeffery, S.F. Madden, P.A. McGettigan, G. Perrière, A.C. Culhane, et D.G. Higgins, "Integrating transcription factor binding site information with gene expression datasets," *Bioinformatics (Oxford, England)*, vol. 23, Fév. 2007, pp. 298-305.
[86] J.H. Phan, Q. Yin-Goen, A.N. Young, et M.D. Wang, "Improving the efficiency of biomarker identification using biological knowledge," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2009, pp. 427-438.
[87] E. Glaab, J.M. Garibaldi, et N. Krasnogor, "ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization," *BMC Bioinformatics*, vol. 10, 2009, p. 358.
[88] Z. Lee, "An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer," *Artificial Intelligence in Medicine*, vol. 42, Jan. 2008, pp. 81-93.
[89] Y. Saeys, T. Abeel, et Y. Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, Antwerp, Belgium: Springer-Verlag, 2008, pp. 313-325.
[90] Agilent Technologies, "Two-Color Microarray-Based Gene Expression Analysis (Quick Amp Labeling)," 2007.
[91] Agilent Technologies, *Agilent Feature Extraction Software (v9.5) User Guide*, 2007.
[92] Rosetta Inpharmatics, *Rosetta Resolver User Guide*, 2006.
[93] "Evolutionary Computation Research Team : http://www.shef.ac.uk/acse/research/ecrg/getgat.html."
[94] S. Shapiro et M. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 3, 1965.
[95] I.T. Young, "Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources," *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society*, vol. 25, Jul. 1977, pp. 935-941.
[96] R.A. Fisher, "The Logic of Inductive Inference," *Journal of the Royal Statistical Society*, vol. 98, 1935, pp. 39-82.

[97] P. Dagnelie, *Théorie et méthodes statistiques : applications agronomiques*, Duculot, Grembloux, 1970.