N° ordre : 2013-16
N° série : G-11

THÈSE / AGROCAMPUS OUEST

Sous le label de l'Université Européenne de Bretagne

pour obtenir le diplôme de :

DOCTEUR DE L'INSTITUT SUPÉRIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE

Spécialité : Mathématiques Appliquées

École doctorale : MATISSE

présentée par :

Marie Verbanck

Analyse exploratoire de données transcriptomiques : de leur visualisation à l'intégration d'information extérieure

Soutenue le **04 septembre 2013** devant la commission d'Examen :

Hervé Abdi	University of Texas at Dallas (USA)	Rapporteur
Philippe Besse	Institut de Mathématiques de Toulouse (France)	Rapporteur
Jean Mosser	CNRS Université de Rennes 1 (France)	Président
Sandrine Lagarrigue	INRA/Agrocampus Ouest, Rennes (France)	Examinatrice
Jérôme Pagès	Agrocampus Ouest, Rennes (France)	Directeur de thèse
Sébastien Lê	Agrocampus Ouest, Rennes (France)	Directeur de thèse

Analyse exploratoire de données transcriptomiques : de leur visualisation à l'intégration d'information extérieure

Ces dernières années ont connu une explosion de recueils de données transcriptomiques à travers l'essor des technologies d'investigation à haut débit. Dans ce travail de recherche, nous proposons de nouvelles méthodologies statistiques exploratoires dédiées au traitement des données transcriptomiques (données de type puce à ADN).
Les données transcriptomiques offrent une image du transcriptome qui lui-même est le résultat des phénomènes d'activation ou d'inhibition de l'expression des gènes. Dans ce cadre, à travers l'analyse statistique des données transcriptomiques, nous cherchons à émettre des hypothèses sur l'expression des gènes ainsi que sur leurs interactions.
Cependant, l'image du transcriptome que fournissent les données transcriptomiques est bruitée. C'est pourquoi, nous abordons dans un premier temps la problématique de débruitage des données transcriptomiques dans un cadre de visualisation. Pour cela, nous proposons une version régularisée de l'analyse en composantes principales. Cette version régularisée permet de mieux reconstituer et visualiser le signal sous-jacent de données bruitées.
Par ailleurs, nous pouvons nous demander si la connaissance du seul transcriptome est suffisante pour démêler la complexité des relations entre gènes. C'est pourquoi nous proposons d'intégrer d'autres sources d'information sur les gènes, et ce de façon active, dans l'analyse des données transcriptomiques. Deux grands mécanismes semblent intervenir dans la régulation de l'expression des gènes, les protéines régulatrices (par exemple les facteurs de transcription) et les réseaux de régulation d'une part, la localisation chromosomique et l'architecture du génome d'autre part.
Dans un premier temps, nous nous focalisons sur la régulation par l'intermédiaire de protéines régulatrices ; nous proposons ainsi un algorithme de classification des gènes basé sur l'intégration de connaissances fonctionnelles sur les gènes, fournies par les annotations Gene Ontology. Cet algorithme fournit des clusters de gènes qui sont similaires à la fois du point de vue de l'expression et de leurs annotations fonctionnelles. Les clusters ainsi constitués sont de meilleurs candidats à l'interprétation que les clusters obtenus en considérant uniquement les données d'expression.
Dans un second temps, nous nous intéressons à l'influence de la localisation chromosomique et de l'organisation du génome sur l'expression des gènes. Nous proposons ainsi de relier l'étude des données transcriptomiques à la localisation chromosomique au sein d'une méthodologie développée en collaboration avec des généticiens.

Mots clés : données transcriptomiques, analyse multidimensionnelle, visualisation, régularisation, intégration d'information extérieure, Gene Ontology, localisation chromosomique

Exploratory Analysis of transcriptomic data : from their visualisation to the integration of external information

Recent years have known an outburst of transcriptomic data collections through the rise of high-throughput technologies. In this research work, we propose new methodologies of exploratory statistics which are dedicated to the analysis of transcriptomic data (DNA microarray data).
Transcriptomic data provide an image of the transcriptome which itself is the result of phenomena of activation or inhibition of gene expression. In this framework, through the statistical analysis of transcriptomic data, we try to formulate hypotheses about the expression of genes as well their interactions.
However, the image of the transcriptome which is provided by transcriptomic data is noisy. That is why, firstly we focus on the issue of transcriptomic data denoising, in a visualisation framework. To do so, we propose a regularised version of principal component analysis. This regularised version allows to better estimate and visualise the underlying signal of noisy data.
In addition, we can wonder if the knowledge of only the transcriptome is enough to understand the complexity of relationships between genes. That is why we propose to integrate other sources of information about genes, and in an active way, in the analysis of transcriptomic data. Two major mechanisms seem to be involved in the regulation of gene expression, regulatory proteins (for instance transcription factors) and regulatory networks on the one hand, chromosomal localisation and genome architecture on the other hand.
Firstly, we focus on the regulation of gene expression by regulatory proteins ; we propose then a gene clustering algorithm based on the integration of functional knowledge about genes, which is provided by Gene Ontology annotations. This algorithm provides clusters constituted by genes which have both similar expression profiles and similar functional annotations. The clusters thus constituted are better candidates for interpretation than clusters obtained by considering only the expression data. Secondly, we focus on the influence of chromosomal localisation and genome organisation on gene expression. We propose therefore to link the study of transcriptomic data to chromosomal localisation in a methodology developed in collaboration with geneticists.

Keywords : transcriptomic data, multidimensional analysis, visualisation, regularisation, integration of biological knowledge, Gene Ontology, chromosomal localisation

REMERCIEMENTS

D'aucuns prétendent que le doctorat est un parcours personnel, voire solitaire, une espèce de rite de passage. Effectivement, il arrive parfois ... souvent que l'on se sente seul au cours de son doctorat. Ainsi, il est indispensable de se rappeler à soi-même qu'un doctorat est avant tout une aventure collective qui implique un très grand nombre de personnes qui soutiennent et guident notre travail, mais également des personnes qui, au quotidien, nous portent et nous réconfortent. C'est pourquoi, à travers ces quelques mots, je souhaite adresser des remerciements très sincères et exprimer une gratitude qui ne transparaît pas toujours au quotidien.

Tout d'abord, je remercie très chaleureusement les rapporteurs de cette thèse Philippe Besse et Hervé Abdi à la fois pour l'intérêt qu'ils ont porté à mon travail et pour leur relecture attentive et leurs conseils avisés qui ont permis d'améliorer ce travail de recherche. Un très grand merci également à Sandrine Lagarrigue et Jean Mosser qui ont accepté de compléter ce jury.

Je remercie vivement tous les membres du laboratoire de mathématiques appliquées.

En premier lieu, je pense évidemment à mes deux directeurs de thèse qui m'ont *couvée* avec beaucoup d'attention pendant ces trois années. Je remercie Jérôme pour m'avoir transmis le goût des statistiques, particulièrement de l'analyse des données et pour avoir été très disponible et de bon conseil. Je vous remercie également pour toutes vos petites histoires et anecdotes truculentes qui ont su illuminer le quotidien. J'adresse un immense merci à Sébastien pour tout ce que vous m'avez apporté aussi bien d'un point de vue scientifique que d'un point de vue humain. Je vous remercie d'avoir été toujours encourageant et réconfortant et par dessus-tout je vous suis très reconnaissante de la confiance sans faille que vous avez en moi et qui m'a permis de mener à bien ce doctorat. Je remercie chaleureusement l'inséparable duo de choc constitué par Julie et François pour m'avoir beaucoup appris. Je remercie profondément Karine, mélange de douceur et d'efficacité, qui est indispensable au labo, David qui est extrêmement bienveillant et utilise volontiers ses qualités de *Mentalist* pour consoler les petites peines, Magalie qui est vraiment toute parfaite, Marine qui fut un modèle à suivre en tant que binôme de TD et

Élisabeth qui connaît tous les secrets de l'agro.

Je remercie tous les doctorants avec qui j'ai partagé des joies, des peines, des *très* courts métrages et des parties de saboteurs pendant ces trois années.

Je pense particulièrement à Emeline avec qui j'ai *tricoté* une véritable amitié qui j'espère continuera 1000 fois. Je remercie ensuite quelqu'un dont mon premier peut être rouge, rosé ou blanc, mon second correspond à mon premier multiplié par 5, mon tout est un super camarade de jeux auquel une énigme ne fait pas peur, j'espère que tu te reconnaîtras Vincent ! Je remercie Tâm pour son extrême gentillesse et ses blagues quotidiennes. Je n'oublie pas non plus de remercier Kadar pour son intégrité qui force le respect. Je remercie également les doctorants rencontrés à travers DocAIR et Nicomaque, Charles, Marion, Pef, Yuna et l'homme à tout faire de Nicomaque Joseph. J'ai une pensée nostalgique pour les équipes d'organisation du festival *Sciences en Cour[t]s*, je remercie l'équipe 2012 : Anne-Marie, Chloé, Élise, Émeline et Marine et l'équipe 2013 : Charles, Coraline, Gaëlle, Nico, Sylvain et Yuna pour tous ces moments inoubliables. J'ai volontairement omis Cécile à qui j'adresse un merci tout particulier pour avoir accepté de tenter et à notre grande surprise de réussir l'aventure *StatistiX* !

J'ajoute que j'ai eu la très grande chance et l'immense honneur d'avoir eu tout au long de ma scolarité des enseignants exceptionnels qui m'ont donné le goût d'apprendre et de transmettre. Il était important pour moi d'avoir une pensée pour eux et de leur dire merci.

Je n'aurais su résister pendant ces trois années sans mes *Survoltés* préférés qui partagent volontiers les rires, les confidences et les soirées *vacherèches* ! Merci à vous tous, Alban, Antoine, Chacha, Crystelle, John Lennon, Léna, Nadiya, Romain, Roxane, Sandra et Yann. J'esquisse également un sourire en pensant à la joyeuse troupe de *Frou-Frou les bains* avec qui ce fut un plaisir et une bouffée d'oxygène de jouer ! Enfin, un énorme merci à Nénette et Anne-C pour être si formidables et toujours présentes pour moi.

Même si « En famille, tout se sait mais rien ne se dit ... », je pense qu'il est grand temps pour moi de leur dire merci. Je remercie mes extraordinaires parents, ma maman qui est parfaite et mon papa qui sait tout faire, pour m'épauler dans tout ce que j'entreprends et être les acteurs principaux dans la réussite de chaque chose que j'accomplie. J'adresse un grand merci à quelqu'un d'unique au monde (heureusement), mon *génialissime* frère qui est la personne la plus drôle et débonnaire que je connaisse (j'ai fait exprès d'employer un mot que tu ne connaîtras pas !). Un très grand merci à mes petites tantes qui sont très présentes et toujours réconfortantes. Je pense à Dan qui a toujours un petit potin à raconter, à Mimi qui est un véritable rayon de soleil et j'adresse des remerciements très sincères à ma deuxième maman Marie-Claude qui a été un soutien exemplaire pendant toutes mes études. Je tiens également à remercier la famille Fraboulet de m'avoir accueillie et soutenue.

Enfin, je ne saurais terminer ces remerciements sans te dire merci pour tout Jean-Gab, merci d'être toi, d'être aussi formidable, d'être mon roc, « you is kind you is smart you is important ... ».

J'adresse, pour finir, une spéciale dédicace au Beau Pédro roi du tango !

AVANT-PROPOS

Contexte

Ce travail de thèse a été conduit au sein du Laboratoire de Mathématiques Appliquées d'Agrocampus Ouest sous la direction de Jérôme Pagès et Sébastien Lê. Étant situé au sein d'une école d'agronomie et en contact permanent avec les autres laboratoires agronomiques, le Laboratoire de Mathématiques Appliquées est tourné vers les applications en agronomie et agroalimentaire. Ayant moi-même un parcours tourné vers la biologie, particulièrement la génétique, mais bouleversé par une rencontre saisissante avec les statistiques, cette philosophie de véritablement dédier les statistiques vers les applications biologiques me correspond parfaitement.

Bien que focalisé sur les développements méthodologiques, ce travail de recherche s'inscrit pleinement dans cet esprit pluridisciplinaire et se situe à l'interface entre la génomique et les statistiques. L'esprit de ce travail de recherche est de proposer des réponses méthodologiques, en termes de méthodes statistiques, à des questionnements biologiques. Le manuscrit proposé jongle donc entre des développements méthodologiques et une volonté de traduire les problématiques biologiques en problématiques statistiques.

Ainsi, la génomique a constitué à la fois une source d'inspiration et un domaine d'application privilégié de ce travail de recherche ; cependant le travail méthodologique qui a été produit pourra être étendu à d'autres domaines d'application.

Motivations

Nous avons choisi de centrer ce travail de recherche sur l'étude des données transcriptomiques (données de puce à ADN). Dans le domaine d'application qu'est la génomique, les données transcriptomiques restent un standard ; on compte par exemple 3341 jeux de données de ce type sur Gene Expression Omnibus au 8 juillet 2013. Classiquement, on recueille l'expression des gènes sans faire d'hypothèse préalable ou sans tenir compte

d'un quelconque *a priori* biologique (sans nécessairement sélectionner des gènes) chez des sujets soumis à différentes conditions expérimentales. Au vu de la grande quantité de données collectées et de la complexité des phénomènes mis en jeu, les hypothèses restent difficiles à formuler, ce qui nécessite le développement de méthodologies statistiques adaptées.

Rappelons le protocole classique d'analyse des données transcriptomiques

0. Pré-traitement des données
1. Tests multiples (gène par gène) pour déterminer une liste de gènes différentiellement exprimés en fonction des conditions expérimentales
2. Visualisation de l'ensemble de gènes différentiellement exprimés via des méthodes d'analyse multidimensionnelle
3. Classification de gènes pour obtenir des *clusters* de gènes coexprimés
4. Caractérisation de l'ensemble des gènes différentiellement exprimés, où de chaque cluster de gènes par une liste de fonctions biologiques (tests d'enrichissement)

Structure du manuscrit

Le chapitre 1 du manuscrit présente le contexte du domaine d'application de ce travail, à savoir la génomique.

Notre apport se situe au niveau de l'amélioration du schéma d'analyse classique. Par ailleurs nous proposons un nouveau point de vue sur les données transcriptomiques.

En effet, d'une part, nous proposons de revisiter la problématique de visualisation des données transcriptomiques en tenant compte du bruit qui peut exister dans les données. Cela nous a conduit à proposer une nouvelle méthode de visualisation qui consiste en une version régularisée de l'analyse en composantes principales qui permet de débruiter les données et d'obtenir ainsi une représentation de celles-ci plus fidèle au signal sous-jacent. La problématique de visualisation et de débruitage des données transcriptomiques est traitée dans le chapitre 2 du manuscrit.

D'autre part, nous proposons de revisiter la problématique de classification de gènes en intégrant de l'information biologique. Ainsi, nous avons développé un nouvel algorithme de classification basé sur l'intégration d'annotations Gene Ontology, dans l'analyse de données transcriptomiques. Cet algorithme permet d'obtenir des clusters dont les gènes sont coexprimés et d'avantage biologiquement apparentés. La problématique de classification de gènes et d'intégration d'information biologique est développée dans le chapitre 3 du manuscrit.

Enfin, nous nous sommes intéressés, en collaboration avec les généticiens, à une nouvelle façon de traiter les données transcriptomiques qui consiste à étudier l'expression des gènes en lien avec la localisation chromosomique. La problématique de prise en compte de la localisation chromosomique est discutée dans le chapitre 4 du manuscrit.

Dans la figure suivante, nous proposons de positionner et synthétiser, par rapport aux objectifs et à la stratégie classique (rouge), les apports de ce travail travail de recherche (vert) :

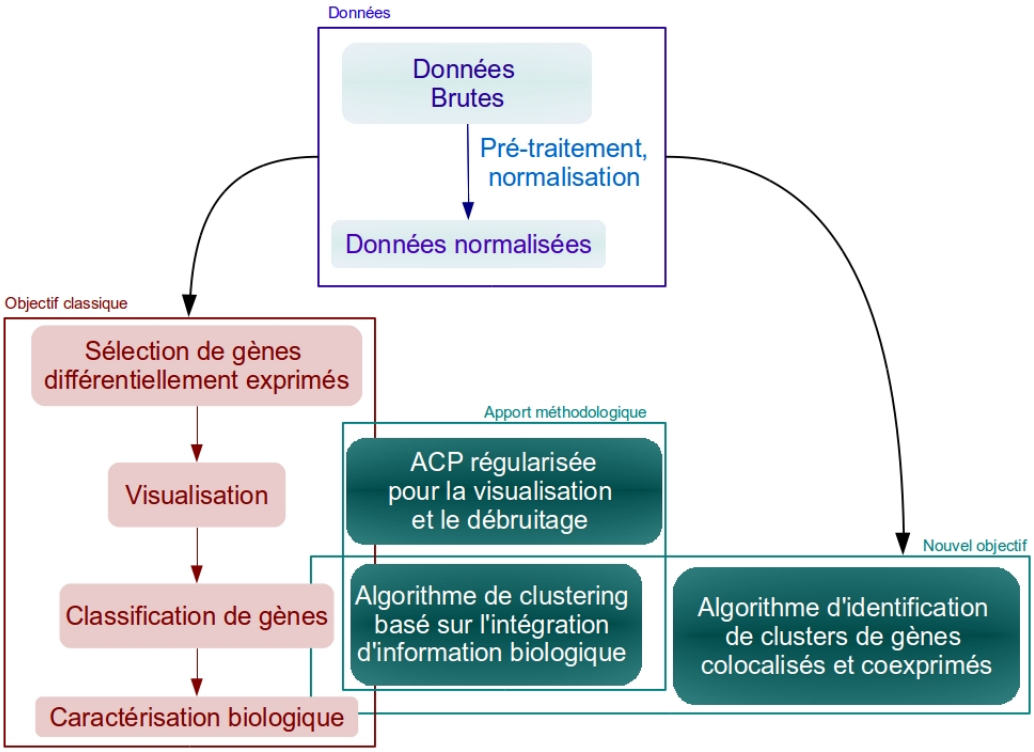


FIGURE 1 – Positionnement du travail de recherche par rapport à la stratégie classique d'étude des données transcriptomiques.

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION

Dans ce chapitre, nous esquissons l'état d'esprit de ce travail de recherche qui exprime véritablement la volonté de partir de la compréhension des phénomènes biologiques, vers la compréhension et la production de méthodologies statistiques adaptées. Une part non négligeable de ce travail a consisté à s'approprier les concepts du domaine d'application afin de les traduire en problématiques statistiques. Nous exposons donc ici les bases de génomique nécessaires au positionnement ainsi qu'à la compréhension de ce travail. Ces considérations biologiques sont mêlées à l'exposé de méthodologies statistiques à la fois classiques et développées au cours de ce travail de recherche.

Sommaire

1 L'expression des gènes : de sa mesure à son interprétation

Pour comprendre les données sur lesquelles nous avons travaillé, nous proposons dans un premier temps de rappeler quelques notions sur l'information génétique et sur l'expression des gènes. Nous nous réservons le droit à quelques simplifications. Précisons que nous nous limiterons à la cellule eucaryote (*i.e.* cellule d'un organisme autre que les bactéries) qui se définit par la présence d'un noyau, isolé du reste de la cellule.

1.1 Organisation de l'information génétique dans la cellule eucaryote

L'information génétique se trouve dans le noyau de la cellule eucaryote et se présente sous forme de chromosomes (schématisé figure 1). Un chromosome est composé de deux molécules d'Acide DésoxyriboNucléique (ADN). Ces deux molécules, communément appelées brins d'ADN, sont constituées d'une séquence ordonnée de nucléotides. Les nucléotides constituent un véritable alphabet du vivant : un nucléotide est caractérisé par une des quatre bases azotées suivantes, A (adénine), T (thymine), C (cytosine) et G (guanine). Leur séquence ordonnée peut constituer des gènes qui spécifient la synthèse d'une

protéine. Les deux brins d'ADN constitutifs du chromosome sont complémentaires, dans le sens où chaque base azotée d'un brin est associée, par des liaisons hydrogènes, à une base azotée de l'autre brin. Cette association se fait selon le motif suivant : A avec T et C avec G. De plus, les deux brins d'ADN présentent une structure dite en double-hélice.

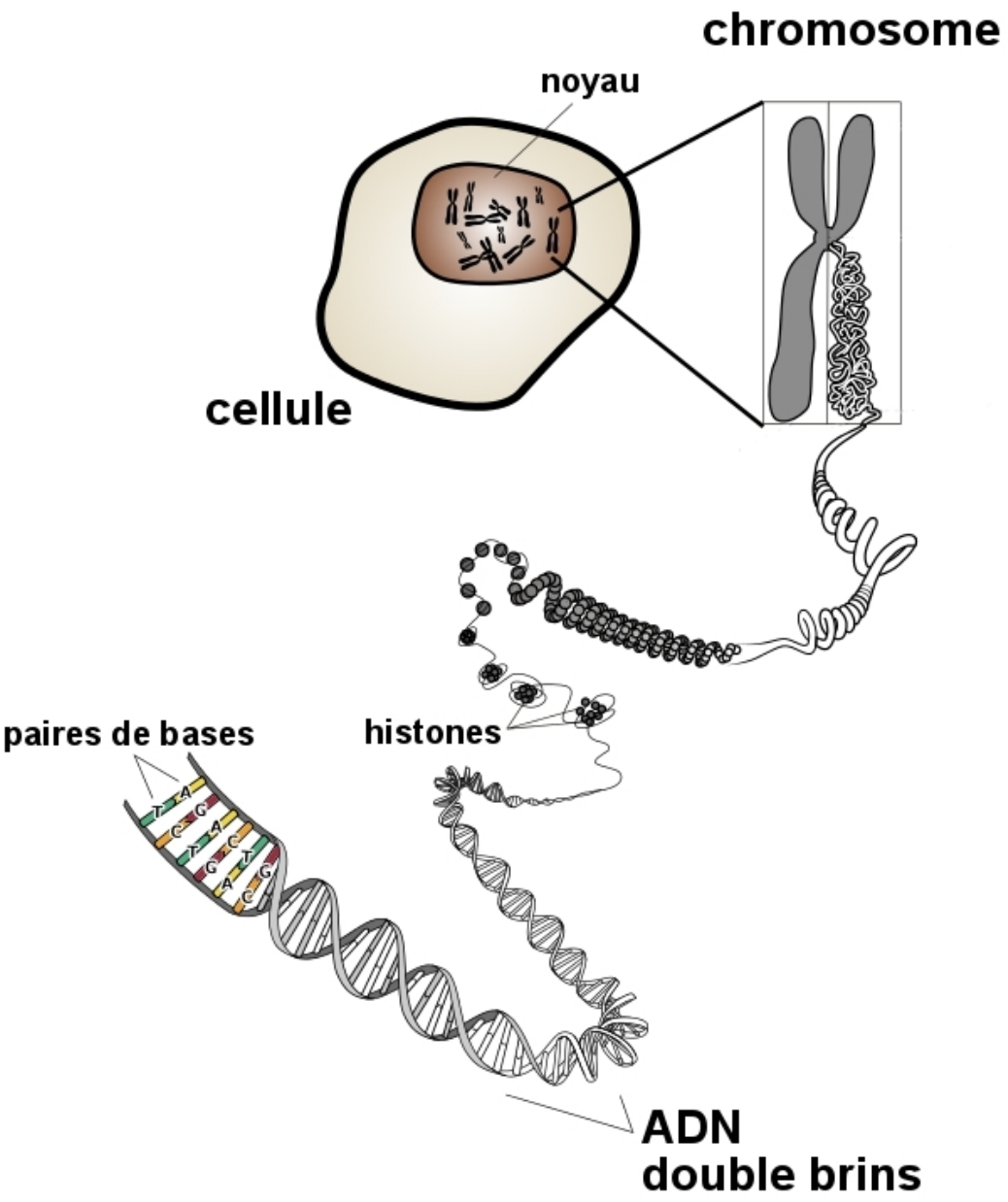


FIGURE 1 – Organisation de l'ADN dans le noyau de la cellule. (*source : http://commons.wikimedia.org*)

Un chromosome n'est pas constitué d'une suite de gènes contigus. Il existe des

séquences non codantes pour des gènes, entre les gènes. De plus, certaines séquences non codantes peuvent être reconnues par des protéines.

Cette information génétique, sous forme de chromosomes, n'est pas libre dans le noyau de la cellule, elle est littéralement « rangée ». Ainsi, les chromosomes sont associés à des complexes protéiques, les histones, qui permettent de les ranger, en les compactant : il faut imaginer chaque chromosome comme un très long filament qui est enroulé successivement et sur toute sa longueur autour des histones (figure 2). Les his-

FIGURE 2 – Schéma de l'enroulement d'un chromosome autour des complexes d'histones. (*source : http://assets.openstudy.com*)

tones peuvent être plus ou moins rapprochées les unes des autres. C'est ce maillage plus ou moins resserré des histones les unes par rapport aux autres qui définit deux états des chromosomes. Un premier état qui correspond à une structure condensée et compacte, dans laquelle les histones sont très resserrées entre elles : c'est l'*hétérochromatine*. Un deuxième état qui correspond à une structure décondensée et lâche, dans laquelle les histones sont peu resserrées entre elles : c'est l'*euchromatine*.

Au sein d'un organisme, l'information génétique est la même d'une cellule à l'autre. Autrement dit, toutes les cellules d'un même organisme possèdent les mêmes gènes. Cependant, tous les gènes ne s'expriment pas dans toutes les cellules : l'expression d'un gène peut dépendre du tissu dans lequel se situe la cellule ou encore de l'environnement de l'organisme.

1.2 L'EXPRESSION DES GÈNES

Pour schématiser, nous pouvons considérer que l'expression d'un gène correspond au mécanisme suivant : à partir d'une séquence d'ADN codant pour un gène, la protéine correspondante est synthétisée. Or, la séquence d'ADN se trouve et reste dans le noyau de la cellule, tandis que la synthèse de la protéine a lieu hors du noyau. Ainsi, l'expression d'un gène nécessite un transfert d'information de l'intérieur du noyau vers l'extérieur, transfert qui se fait par un intermédiaire justement appelé l'Acide RiboNucléique messager (ARNm), pour sa fonction de messager. Ce transfert d'information consiste à synthétiser un ARNm en recopiant la séquence d'ADN codant pour le gène, étape que l'on qualifie de transcription. L'ARNm est ensuite transféré hors du noyau. Enfin, la séquence d'ARNm (copie de séquence d'ADN codant pour le gène) est traduite en une séquence protéique pour fabriquer une protéine, étape que l'on nomme traduction. Le principe de l'expression d'un gène est schématisé figure 3.

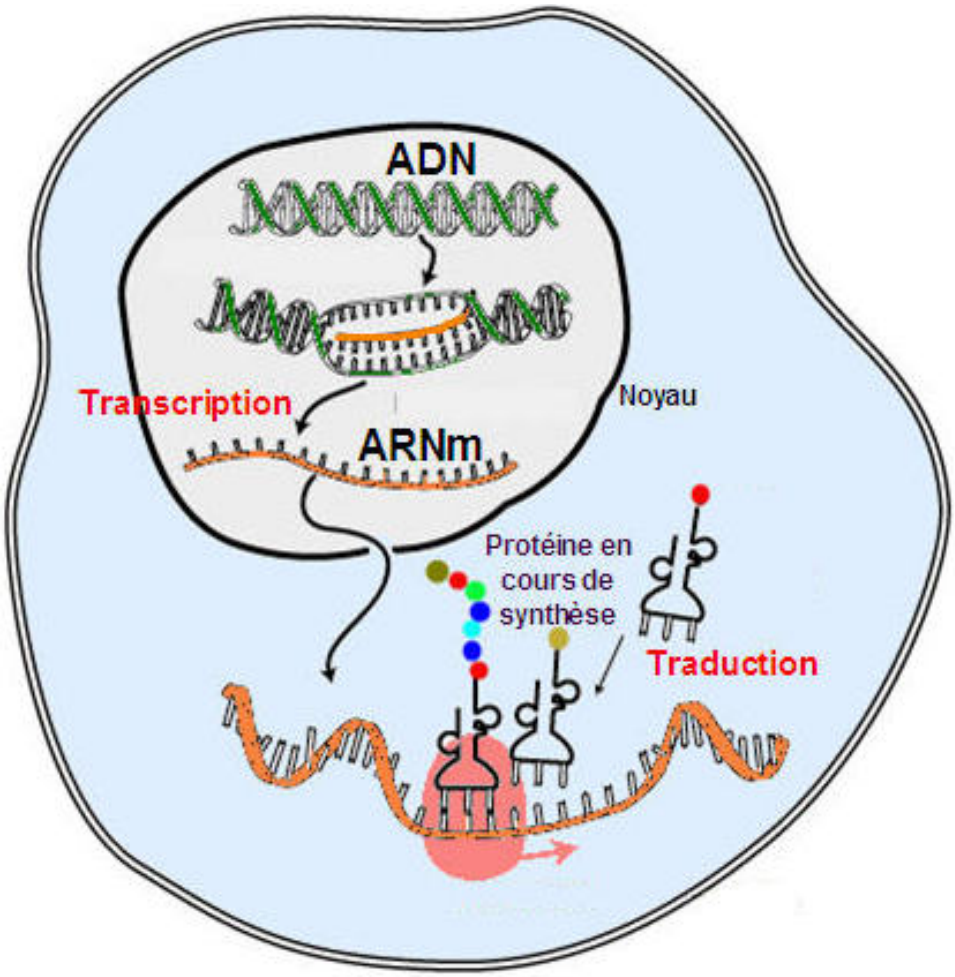


FIGURE 3 – Principe général de l'expression d'un gène. (*source : http://www.linternaute.com/science/biologie/dossiers/06/0609-adn/adn2/2.shtml*)

Plus précisément, la transcription est constituée de plusieurs étapes et fait intervenir une importante machinerie cellulaire. Cette machinerie cellulaire comprend notamment une enzyme, l'ARN polymérase, qui catalyse la synthèse d'ARNm. Rappelons que la synthèse d'ARNm consiste à recopier la séquence d'ADN. Les deux brins d'ADN sont donc séparés et le chromosome est ouvert pour permettre à l'ARN polymérase de synthétiser l'ARNm. La séquence d'ARNm est construite en assemblant des nucléotides, dits précurseurs qui sont libres dans le noyau, par complémentarité avec la séquence d'ADN. Rappelons par ailleurs, que la séquence d'ADN n'est pas constituée d'une suite de gènes contigus, mais qu'il existe des séquences non codantes entre les gènes et que certaines de ces séquences sont des sites de fixation pour des protéines (section 1.1). Or la machinerie cellulaire de la transcription se compose de protéines qui se fixent à l'ADN, les facteurs de transcription, dits généraux. Ainsi sur la séquence d'ADN, en amont du gène à transcrire, se trouve un (ou plusieurs) site de fixation pour des facteurs

de transcription généraux. La fixation des facteurs de transcription est indispensable pour initialiser la transcription. De plus, pour que la transcription ait lieu, il est impératif que la séquence d'ADN à transcrire soit accessible. En effet, rappelons que l'ADN est enroulé autour des histones qui sont ensuite resserrées les unes des autres (section 1.1). Ainsi le gène à transcrire doit se trouver sous forme d'euchromatine (maillage des histones lâche) et des protéines interviennent pour modifier légèrement l'enroulement du chromosome autour des histones, ce qui rend accessible la séquence d'ADN à transcrire. [1] [2]

Le transcriptome est défini comme l'ensemble des ARNm présents dans un type cellulaire donné, à un temps donné et dans une condition biologique ou expérimentale donnée.

1.3 Données transcriptomiques (puce à ADN)

Le recueil de données transcriptomiques consiste en une quantification du **transcriptome**. Cette quantification consiste à estimer le nombre de copies d'ARNm pour chaque gène dans un tissu cellulaire donné, à un temps donné et dans une condition donnée. La puce à ADN est un des outils permettant la quantification du transcriptome d'un sujet. Le principe de la puce à ADN s'appuie sur la propriété d'hybridation de l'ADN (section 1.1), qui repose sur les liaisons hydrogènes qui existent entre les bases azotées. [3] Rappelons succinctement la structure de la puce à ADN et le principe du recueil des données transcriptomiques (figure 4).

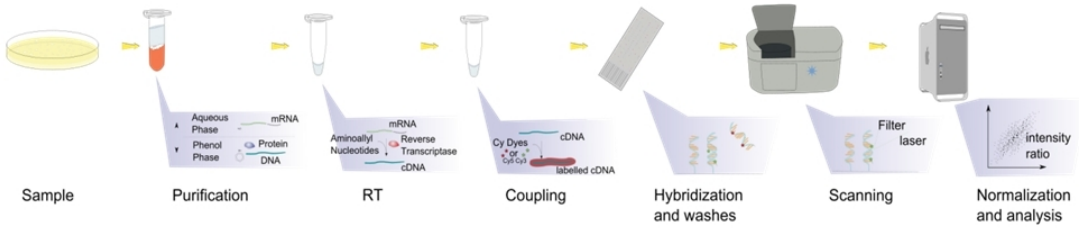


FIGURE 4 – Principe du recueil de données transcriptomiques. (*source : http://en.wikipedia.org/wiki/DNA_microarray*)

1. Les ARMm subissent un certain nombre de modifications post-transcriptionnelles que nous ne détaillerons pas dans un soucis de simplification.
2. Suite à la transcription, les ARNm sont exportés hors du noyau de la cellule pour être traduits en protéines. Cette étape, que nous ne détaillerons pas ici, est la traduction.
3. Doser l'ARNm est plus facile que de doser les protéines, et la puce à ADN permet un dosage haut-débit, ce qui n'est pas possible pour les protéines. Cependant, sachant que l'ARNm n'est qu'un intermédiaire entre la séquence d'ADN et la protéine, on s'attend à ce que les quantités d'ARNm reflètent les quantités de protéines correspondantes.

1.3.1 La puce à ADN

Une puce à ADN est un support solide constitué de puits. Dans chaque puits, sont fixées des séquences d'ADN simple brin spécifiques d'un seul et unique gène. Les ARNm du sujet à étudier sont extraits et purifiés. Puis à partir de ces ARNm, sont synthétisées des molécules d'ADN simple brin cycliques (ADNc) qui sont complémentaires des ARNm du sujet. Chaque molécule d'ADNc est ensuite couplée à une molécule fluorescente. Cette molécule fluorescente n'est qu'un intermédiaire technique qui permettra la mesure. L'ensemble des ADNc est hybridé sur la puce à ADN : les ADNc se fixent aux séquences des puits qui leur sont complémentaires. Le principe de l'hybridation repose sur les liaisons hydrogènes qui existent entre les bases azotées complémentaires, c'est ce principe qui permet la structure double brins de l'ADN de la cellule (figure 5). La puce est ensuite lavée

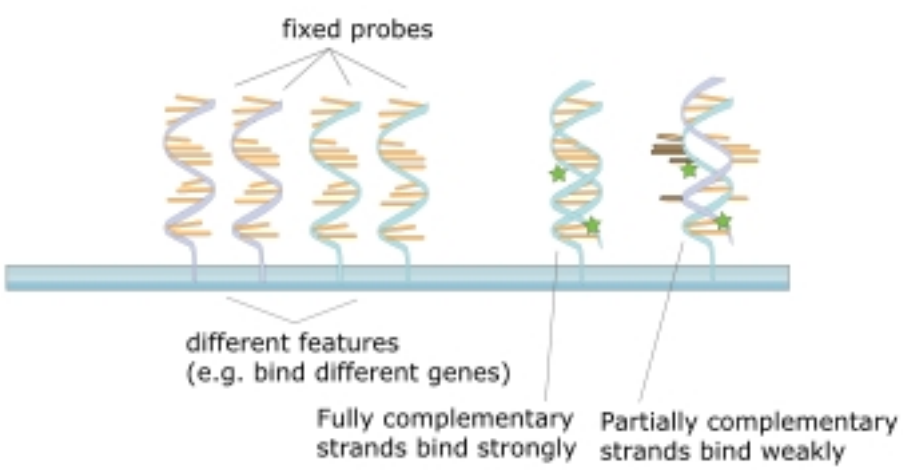


FIGURE 5 – Principe de l'hybridation. A gauche, quatre séquences d'ADN simple brin qui sont fixées dans le puits. A droite, deux séquences d'ADNc hybridées, l'une entièrement hybridée (gauche), l'autre partiellement (droite). (*source : http://en.wikipedia.org/wiki/DNA_microarray*)

pour éliminer les ADNc qui ne sont pas hybridés ou qui sont partiellement hybridés (si la séquence n'est pas parfaitement complémentaire par exemple). Seuls les ADNc hybridés demeurent sur la puce.

Un laser est ensuite utilisé pour exciter les molécules fluorescentes attachées à chaque ADNc : la quantité de fluorescence émise donne une estimation du nombre d'ADNc hybridés dans chaque puits. La donnée brute issue d'un tel recueil de données est une image. C'est cette image qui est analysée et qui permet de quantifier, en fonction du niveau de fluorescence, la quantité d'ADNc hybridé dans chaque puits. La quantité d'ADNc hybridé est donc proportionnelle à la quantité d'ARNm extraite des cellules du sujet. L'exploitation des images de fluorescence n'est pas immédiate et nécessite un pré-traitement.

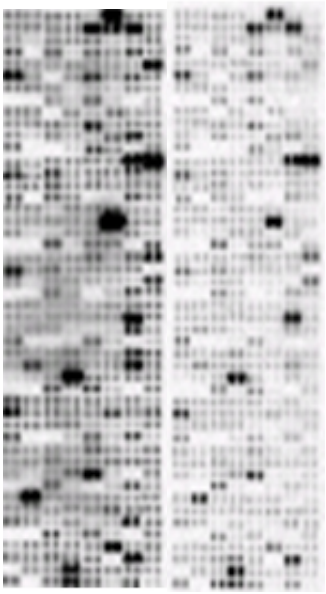
1.3.2 Pré-traitement des données transcriptomiques

La première transformation des données de fluorescence brute est le passage au log2. Cela permet de se rapprocher d'une distribution gaussienne. Par ailleurs, il est possible de considérer pour chaque gène le ratio d'expression entre une condition donnée et une condition de référence. Dans ce cas le passage au logarithme permet de traiter les ratios et leur inverse de façon symétrique (par exemple $\log_2(4) = 2$, $\log_2(1/4) = -2$).

La comparaison directe des données brutes est délicate du fait de plusieurs sources de variabilité technologiques (Quackenbush, 2002). L'étape de normalisation consiste à repérer puis à corriger si possible les biais systématiques dans la mesure de l'expression. Les principales sources de variations que nous pouvons identifier sont les suivantes :

- la quantité d'ADNc qui est déposée sur la puce peut varier d'une puce à l'autre
- en plus des ADNc en surplus, le lavage de la puce peut être plus ou moins stringent et retirer des ADNc hybridés
- il peut exister des différences dans le couplage des ADNc aux molécules fluorescentes (association d'aucune ou de plus d'une molécule à certains ADNc)

Ces différentes sources de variabilité conduisent à des différences globales et systématiques entre puces. Si nous prenons l'exemple de ces deux puces (figure 6), la puce de gauche (quel que soit le gène) présente systématiquement des valeurs d'expression plus élevées. Ceci est uniquement dû aux biais technologiques précédemment énoncés.



(a) (b)

FIGURE 6 – Exemple de deux puces.

Il est par conséquent naturel de considérer un effet puce et d'associer à chaque puce un coefficient correcteur global. Nous construisons un modèle d'analyse de variance par

gène :

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r + \varepsilon_{ir}^{(k)} \quad (1)$$

avec $Y_{ir}^{(k)}$ l'expression du gène k sur la $r^{\text{ème}}$ puce de la condition expérimentale i, $\mu^{(k)}$ l'expression moyenne du gène k sur toutes les puces, $\alpha_i^{(k)}$ l'effet condition expérimentale, égale à l'expression moyenne du gène k chez toutes les puces de la condition i, β_r l'effet puce qui ne dépend pas du gène k (ajusté toutes conditions confondues) qui est associé à chaque puce et correspond au coefficient correcteur global de chaque puce.

Pour ajuster cet effet puce et donc calculer un coefficient correcteur global par puce, il est nécessaire de se fonder sur des hypothèses biologiques :

- « La majorité des gènes n'est pas différentiellement exprimée/régulée entre les différentes conditions expérimentales », « la régulation totale du génome est relativement constante ». Autrement dit, si nous comparons les individus d'une condition expérimentale à l'autre, seule une petite proportion de gènes (étant donné le nombre de gènes) s'exprime différemment (assomption fausse pour les puces de gènes présélectionnés)
- Le nombre moyen de gènes qui s'expriment est le même d'une condition expérimentale à l'autre

Ces deux hypothèses conduisent à penser que la moyenne des expressions sur une puce est la même d'une puce à l'autre. Ainsi, les moyennes globales d'expression devant être constantes, il est possible d'estimer l'effet puce. Une fois cet effet puce estimé, nous obtenons un coefficient correcteur spécifique de chaque puce (β_r) qui permet de corriger l'expression de l'ensemble des gènes de la puce.

Remarque : Il existe des technologies de puce à ADN bicouleur, que nous ne détaillerons pas, où sur une même puce sont hybridés les ADNc issus de deux sujets différents qui ont au préalable été couplés à des molécules fluorescentes de deux couleurs différentes.

1.4 Démarche statistique classique d'analyse des données transcriptomiques

Il existe un grand nombre d'outils statistiques dédiés à l'analyse de données transcriptomiques. Afin de positionner ce travail de recherche, nous proposons de présenter succinctement les principales méthodologies utilisées selon un schéma classique (pour une revue complète voir Brazma et Culhane (2004)).

1.4.1 Détection de gènes différentiellement exprimés

Les données transcriptomiques permettent de mesurer l'expression d'un grand nombre de gènes sans *a priori*, potentiellement tous les gènes connus pour s'exprimer dans un tissu par exemple. Cependant, seuls les gènes dont l'expression varie entre les différentes conditions expérimentales présentent un intérêt pour le biologiste. Ainsi, la détection de ces gènes dits différentiellement exprimés est primordiale et constitue une étape préalable de l'analyse des données transcriptomiques. Les tests multiples sont

très largement utilisés pour détecter cet ensemble de gènes différentiellement exprimés. Classiquement, les tests multiples consistent à ajuster un modèle par expression de gène, en fonction des différentes conditions expérimentales, ce qui fournit une probabilité critique par gène. Plusieurs solutions de tests multiples existent pour fixer la valeur du seuil de significativité des probabilités critiques. D'autres méthodes ajustent un modèle par gène légèrement différent en intégrant par exemple des facteurs prenant en compte la dépendance des données (Friguet *et al.*, 2009). Les tests multiples pour l'analyse de données transcriptomiques sont donc un domaine de recherche actif (pour une revue des méthodes communes de tests multiples voir Jeanmougin *et al.* (2010)).

Remarque : la détection de gènes différentiellement exprimés est une procédure de sélection de gènes. Nous pouvons donc citer des méthodes d'analyse multidimensionnelle dites parcimonieuses qui se sont également développées récemment. En appliquant par exemple des contraintes de parcimonie sur les loadings d'une analyse en composantes principales sur le tableau d'expression, les sujets étant les individus et les gènes les variables, on sélectionne des gènes (Witten et al.*, 2009).*

1.4.2 Visualisation des données

La visualisation des données est un élément majeur de l'analyse des données transcriptomiques. Les méthodes les plus communément utilisées sont des méthodes d'analyses factorielles, telles que l'analyse en composantes principales (ACP). L'analyse exploratoire étant au centre de ce travail de recherche, un bilan des méthodes de visualisation dans le cadre de l'analyse des données transcriptomiques sera présenté dans le chapitre 2 de ce manuscrit.

1.4.3 Clustering de gènes

La classification ou plus génériquement le *clustering* de gènes est une étape indispensable à l'interprétation des données puisque c'est fréquemment sur la base des groupes, encore appelés *clusters* ou *modules*, de gènes que se fait l'interprétation et la caractérisation biologique.[4] Voici un aperçu, non exhaustif, des méthodes de clustering usuelles.

La méthode de clustering, probablement la plus repandue, consiste à appliquer un algorithme de classification ascendante hiérarchique sur une matrice de distance construite à partir des coefficients de corrélations entre expressions géniques. La classification ascendante hiérarchique fournit un dendrogramme qui doit être coupé à une certaine hauteur pour founir une partition. La variante la plus répandue de cet algorithme est le Heatmap (Eisen *et al.*, 1998). Il consiste en une double classification ascendante hiérarchique à la fois sur les gènes et sur les sujets. De plus, on trouve également des versions où l'algorithme de clustering est un algorithme K-means qui implique un choix *a priori* du nombre de classes (Hartigan et Wong, 1979). Notons qu'il existe des variantes descendantes de

4. Par la suite, nous nous permettrons d'utiliser les anglicismes *clustering* et *clusters*.

clustering, par exemple l'algorithme Diana (Kaufman et Rousseeuw, 1990). Parmi les techniques de clustering, nous pouvons également citer les méthodes de clustering basées sur des modèles. Les données transcriptomiques sont par exemple modélisées par des lois de mélange gaussiennes, où chaque loi représente un cluster. Le maximum de vraisemblance (via un algorithme EM) est utilisé pour ajuster les paramètres des distributions et déterminer l'appartenance des gènes aux différents clusters (Banfield et Raftery, 1993).

1.4.4 Caractérisation biologique

La caractérisation biologique est une étape clé de l'analyse des données transcriptomiques puisque c'est l'étape qui permet l'interprétation biologique à proprement parler. Cette étape nécessite l'utilisation d'information biologique extérieure sur les gènes, telle que des annotations fonctionnelles de type Gene Ontology (Ashburner *et al.*, 2000), et consiste à attribuer à un ensemble de gènes (issu soit d'une sélection, soit d'un clustering), des fonctions biologiques qui le caractérisent. Classiquement, déterminer si une fonction caractérise ou non un ensemble revient à effectuer un test exact de Fisher qui vise à comparer la proportion de gènes associés à la fonction dans l'ensemble à caractériser par rapport à la proportion de gènes associés à la fonction dans un ensemble de référence (Fisher, 1922).

2 Amélioration de l'analyse des données transcriptomiques

Ce travail de recherche est centré autour de l'amélioration de l'analyse des données transcriptomiques à travers deux points de vue. D'une part, nous pouvons penser que les données transcriptomiques sont bruitées, ce qui nous amène à mettre au point une stratégie de débruitage. D'autre part, nous pouvons penser que les données transcriptomiques ne se suffisent pas à elles-mêmes, c'est pourquoi nous proposons des stratégies d'intégration d'information extérieure.

2.1 Visualisation et débruitage des données transcriptomiques

Rappelons que les données transcriptomiques fournissent une image du transcriptome. C'est à partir de cette image que l'on tire des hypothèses sur l'expression des gènes et particulièrement sur les relations qui existent entre l'expression de plusieurs gènes. Néanmoins, cette image peut être faussée, ce qui se traduit par des données bruitées qui conduisent à une mauvaise exploitation des données.

Précisons cependant, que dans le pré-traitement des données transcriptomiques, le débruitage des données, notamment le débruitage des images brutes est un enjeu majeur du pré-traitement. Cependant, il reste nécessairement une incertitude sur la mesure.

Ainsi, les données transcriptomiques peuvent être vues comme un « véritable » signal entaché d'erreur. Nous considérerons donc un modèle signal + bruit afin de proposer une méthodologie de débruitage des données transcriptomiques dans un cadre de visualisation, autrement dit dans un cadre exploratoire. Cet axe de travail est développé dans **le chapitre 2 de ce manuscrit**. Nous proposons tout d'abord de faire un état de l'art des méthodes et pratiques de visualisation des données transcriptomiques. Nous proposons ensuite de rappeler l'intérêt de ces outils de visualisation. Enfin la question du débruitage des données transcriptomiques nous amènera à proposer une nouvelle méthode de visualisation des données transcriptomiques qui prend en compte la structure bruitée des données. Ainsi, nous proposons une version régularisée de l'analyse en composantes principales. Cette version régularisée permet de mieux reconstituer et visualiser le signal sous-jacent de données bruitées.

2.2 Intégration d'information extérieure

Rappelons que toutes les cellules d'un organisme eucaryote possèdent les mêmes chromosomes et renferment donc exactement la même information génétique (section 1.1). Or, tous les gènes ne s'expriment pas dans toutes les cellules, ce que l'on peut observer entre deux cellules issues de deux tissus différents par exemple. Ainsi la même information génétique est utilisée de façon différente selon le type de cellule mais également en fonction de l'environnement. Il existe donc une multitude de phénomènes de régulation de l'expression des gènes.

Le transcriptome, que l'on mesure, est donc le résultat de ces phénomènes de régulation de l'expression des gènes. Nous pouvons supposer que différents mécanismes de régulation de l'expression (plus précisément de la transcription sachant que nous nous intéressons aux ARNm) sont mis en place entre les différentes conditions expérimentales. C'est donc à partir de la comparaison des transcriptomes entre les différentes conditions expérimentales que nous cherchons à comprendre les phénomènes de régulation et d'interaction entre gènes.

Précisons que pour interpréter les données transcriptomiques, nous utilisons deux règles simples. D'une part, nous apportons des jugements relatifs en fonction des conditions expérimentales, nous dirons qu'un gène est sur-exprimé ou sous-exprimés dans une condition expérimentale par rapport à une autre condition. D'autre part, nous définissons la notion de coexpression de deux gènes. Si deux gènes sont sur-exprimés dans une même condition et sous-exprimés dans une même autre condition expérimentale, autrement dit si les profils d'expression de ces deux gènes sont corrélés positivement, nous qualifierons ces gènes de coexprimés. Précisons que dans certains cas la notion de coexpression pourra s'étendre aux gènes dont les profils d'expression sont fortement corrélés négativement.

Prenons un exemple concret d'interprétation. Considérons deux gènes qui sont tous deux sur-exprimés chez les sujets soumis à la condition expérimentale numéro 1 et sous-exprimés chez les sujets soumis à la condition expérimentale numéro 2 : ces deux gènes

sont donc coexprimés. Nous pouvons penser que l'expression de ces deux gènes a été activée et/ou inhibée dans une condition par rapport à l'autre, par le même agent régulateur. Nous pouvons encore penser que l'un des deux gènes active l'expression du second, ainsi si le gène activateur s'exprime, l'autre s'exprime et *vice versa*. Enfin, nous pouvons penser que l'expression de ces deux gènes a été activée/inhibée par deux agents indépendants mais dont la synthèse a été activée en parallèle au même moment.

Ainsi, la coexpression des gènes peut conduire à une large variété d'interprétations. Or, la stratégie classique d'analyse des données transcriptomiques repose *in fine* sur le regroupement des gènes en fonction de leur coexpression, ce qui fournit des clusters de gènes coexprimés qui sont ensuite interprétés (section 1.4). Par conséquent, nous pouvons penser que le clustering de gènes basé uniquement sur la coexpression n'est pas suffisant pour démêler les relations complexes qui existent entre les gènes et recouvrer tous les phénomènes de régulation de l'expression des gènes.

C'est pourquoi, nous proposons d'intégrer d'autres sources d'information extérieure sur les gènes dans un cadre de clustering. Nous proposons de prendre en compte deux types d'information extérieure sur les gènes ce qui a donné lieu à deux travaux différents. D'une part, nous proposons d'intégrer des annotations fonctionnelles sur les gènes, ce qui a constitué une grande partie de ce travail de recherche. D'autre part, à l'occasion d'un travail collaboratif qui a été effectué avec le laboratoire de génétique animale d'Agrocampus Ouest dans le cadre du doctorat de Marion Ouédraogo, nous nous sommes intéressés à la prise en compte d'une information de localisation chromosomique des gènes.

2.2.1 Intégration d'information de type Gene Ontology

Afin de comprendre comment nous avons abouti à la mise au point d'une méthodologie d'intégration d'information de type Gene Ontology (Ashburner *et al.*, 2000) qui rassemble des annotations sur le rôle des gènes, nous proposons de présenter notre approche depuis la connaissance théorique des phénomènes de régulation de l'expression des gènes vers l'utilisation concrète de l'information disponible.

Comme nous l'avons évoqué, tirer des conclusions sur les phénomènes de régulation de la transcription à partir de la comparaison des transcriptomes est délicat. C'est pourquoi nous sommes partis de la connaissance théorique des phénomènes de régulation de la transcription que nous rappelons.

L'ADN des cellules eucaryotes est constitué de séquences codantes (gènes) mais majoritairement de séquences non codantes. Parmi ces séquences non codantes se trouvent des sites de fixation à l'ADN pour des protéines (section 1.1). En effet, nous avons déjà évoqué la fixation de facteurs de transcription généraux à des séquences spécifiques d'ADN en amont du gène qui jouent un rôle prépondérant dans la transcription (section 1.2). En plus de ces facteurs de transcription généraux, il existe un grand nombre de protéines régulatrices plus ou moins spécifiques (*i.e.* pour un seul ou plusieurs gènes) qui jouent un rôle majeur dans la régulation de la transcription. Ainsi, un des grands mécanismes de régulation de l'expression des gènes est permis par l'action combinée de

plusieurs protéines régulatrices. Les protéines régulatrices peuvent soit activer, soit inhiber la transcription, cependant les phénomènes d'activation semblent être prépondérants chez les eucaryotes. Or une protéine régulatrice, comme son nom l'indique, est bien une protéine qui provient nécessairement de l'expression d'un autre gène. On retrouve, par conséquent, beaucoup de phénomènes de cascades de réactions du type « un gène dont la protéine active l'expression d'un autre gène dont la protéine active l'expression d'un autre gène et ainsi de suite ». Ce principe est la base du réseau de régulation génique.

Ainsi, des gènes qui se régulent entre eux font partie d'un même processus biologique. Or nous disposons d'information biologique sur les gènes, comme la base de données Gene Ontology (Ashburner *et al.*, 2000) qui tente de rassembler toutes les informations disponibles sur les gènes sous forme d'annotations fonctionnelles. Ces annotations fonctionnelles décrivent notamment l'appartenance des gènes à des processus biologiques.

Le chapitre 3 de ce manuscrit est centré sur l'intégration de connaissances biologiques synthétisées par des annotations fonctionnelles de type Gene Ontology. Les connaissances fonctionnelles sur les gènes sont utilisées dans l'esprit suivant : considérons deux gènes qui sont coexprimés, si ces deux gènes partagent un grand nombre d'annotations fonctionnelles, nous pouvons penser qu'il existe un véritable lien de régulation entre ces deux gènes, autrement dit l'expression de ces deux gènes est peut-être activée par une même protéine régulatrice. Nous proposons un algorithme de clustering qui fournit des clusters de gènes similaires à la fois du point de vue de l'expression et de leurs annotations fonctionnelles. Les clusters ainsi constitués sont de meilleurs candidats à l'interprétation.

2.2.2 Prise en compte de la localisation chromosomique

Des études ont montré qu'il existe des éléments régulateurs qui contrôlent l'expression de plusieurs gènes contigus sur les chromosomes. Certains éléments régulateurs peuvent se fixer en amont d'un groupe de gènes contigus et activer la transcription de l'ensemble des gènes de la région (Blumenthal, 2004). A l'inverse, d'autres types d'éléments régulateurs peuvent se fixer en aval d'une région et permettent ainsi la formation de boucles dans le chromosome ce qui résulte en l'activation de l'expression de certains gènes et en l'inhibition de l'expression d'autres gènes de la région (Dean, 2011).

Par ailleurs, des théories récentes ont proposé que la localisation des gènes dans le noyau de la cellule a un rôle majeur dans la régulation de leur expression. En premier lieu, la structure de l'ADN est plus ou moins compactée permettant l'accessibilité ou non de la machine transcriptionnelle aux gènes à transcrire.

Les premiers indices de ce phénomène de compaction ont été montrés par microscopie électronique (figure 7). Nous distinguons des territoires clairs (1) qui correspondent à une structure décondensée des chromosomes : l'euchromatine et des territoires foncés (2) qui correspondent à une structure condensée : l'hétérochomatine (section 1.1) [5]. Cette étude a montré que l'euchromatine correspond à un état « actif » de l'ADN puisque cette

5. Le territoire foncé au centre du noyau, noté N, est le nucléole dont nous ne parlerons pas ici.

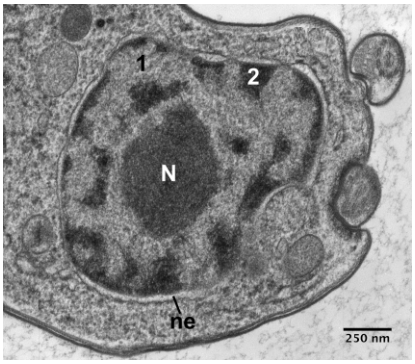


FIGURE 7 – Coupe d'un noyau de cellule (ne : membrane du noyau, 1 : euchromatine, 2 : hétérochromatine, N : nucléole) (Daniels *et al.*, 2010).

structure permet l'expression des gènes, tandis que l'hétérochromatine correspond à un état « inactif » de l'ADN puisque cette conformation ne permet pas l'expression des gènes, les gènes sont dits éteints. Nous remarquons donc que l'organisation dans le noyau ne semble pas due au hasard et nous distinguons véritablement l'euchromatine qui est située au centre du noyau (autour du nucléole) et l'hétérochromatine à la périphérie.

De plus, des études ont permis de marquer les chromosomes afin de mettre en évidence leur position dans le noyau de la cellule. Ces études ont pu montrer que chaque chromosome a une localisation précise et occupe un espace distinct dans le noyau (Croft *et al.*, 1999) : on parle de territoires chromosomiques. Ainsi les chromosomes comportant le plus de gènes actifs sont sous forme d'euchromatine et se trouvent au centre du noyau tandis que les chromosomes comportant moins de gènes actifs sont sous forme d'hétérochromatine et se situent à la périphérie du noyau. S'ajoute à l'organisation en territoires chromosomiques, le fait qu'à l'interface de ces territoires aient lieu les événements de transcription (Zirbel *et al.*, 1993). Le concept d'usine de transcription a été introduit pour décrire ces interfaces des territoires chromosomiques dans lesquelles les chromosomes s'entremêlent et interagissent.

Ainsi, ces considérations nous permettent de penser que la localisation chromosomique des gènes ainsi que les interactions entre chromosomes ont une grande importance dans la régulation de l'expression des gènes. Nous avons donc développé, en collaboration avec les généticiens du laboratoire de génétique animale d'Agrocampus Ouest à l'occasion du travail de doctorat de Marion Ouédraogo, une méthodologie d'étude des données transcriptomiques en lien avec la localisation chromosomique. Ce travail collaboratif est présenté dans **le chapitre 4**.

3 Jeux de données illustratifs

Nous présentons ici deux jeux de données illustratifs qui seront utilisés dans les différents chapitres. Ces jeux de données seront notamment utilisés dans **le chapitre 5**

qui présente les différents développements logiciels issus de ce travail de recherche.

Nous avons utilisé un jeu de données transcriptomiques généré suite à une problématique agronomique et issu d'un jeu de données publié dans Désert *et al.* (2008), accessible sous l'identifiant GSE11290 sur la plateforme Gene Expression Omnibus (Barrett et Edgar, 2006). Dans la publication originale, les poulets étaient soumis à 3 conditions expérimentales : jeûne de 16 heures (J16), jeûne de 48 heures (J48) et nourris en continu (N). Nous avons conservé les conditions J16 et N en rajoutant deux conditions : jeûne de 16 heures puis renutrition de 5 heures (J16R5), jeûne de 16 heures puis renutrition de 16 heures (J16R16) car de nouvelles données ont été générées au laboratoire de génétique d'Agrocampus Ouest. *In fine*, nous avons travaillé sur un jeu de données comptant 27 poulets soumis à 4 conditions expérimentales : J16, J16R5, J16R16 et N.

Par ailleurs, nous avons utilisé un jeu de données qui n'est pas un jeu de données transcriptomiques mais un recueil d'images. Il s'agit du jeu de données PINCAT (Sharif et Bresler, 2007) qui rassemble des images d'IRM en temps réel et le jeu de données se compose de 50 images. Les données sont donc sous forme d'un cube de données avec sur les deux premières dimensions la largeur et la hauteur de l'image et à l'intersection la valeur pour un pixel, et la troisième dimension représente les différents temps qui sont au nombre de 50 dans cet exemple.

CHAPITRE 2

VISUALISATION ET DÉBRUITAGE DES DONNÉES TRANSCRIPTOMIQUES

Dans ce chapitre, nous proposons une nouvelle méthodologie de débruitage de données dans un cadre de visualisation sous la forme d'une analyse en composantes principales régularisée. Au préalable, nous faisons un bilan des pratiques de l'application de l'analyse en composantes principales aux données transcriptomiques. Il s'avère que l'analyse en composantes principales est particulièrement utilisée pour débruiter les données transcriptomiques. En effet, les données transcriptomiques peuvent être considérées comme le mélange d'un signal d'intérêt mais que l'on ne connaît pas et de bruit. C'est pourquoi nous proposons une version régularisée de l'analyse en composantes principales qui est une méthode prometteuse pour débruiter les données dans un cadre de visualisation.

Ce chapitre inclut l'article :

Verbanck, M., Josse, J., & Husson, F. (2013). Regularised PCA to visualise and denoise data. *Statistics and Computing* (soumis)

Sommaire

Dans le traitement des données transcriptomiques, l'analyse en composantes principales (ACP) est utilisée à la fois comme un outil de visualisation, mais également comme un outil de pré-traitement des données.

Dans un premier temps, nous proposons de nous focaliser sur les pratiques et l'utilisation de l'ACP comme un outil de visualisation des données transcriptomiques. Ce bilan nous conduira à réaffirmer les liens entre l'ACP et les données transcriptomiques qui ne

sont pas clairement établis et qui peuvent de fait conduire à un appauvrissement de l'interprétation des ACP appliquées à ce type de données que l'on trouve dans la littérature.

Dans un deuxième temps, sachant l'ACP est très utilisée comme outil de prétraitement des données transcriptomiques, nous proposons également de dresser un bilan des ces pratiques. Nous verrons que dans cette utilisation, l'ACP consiste à estimer une matrice de données à partir de la formule de reconstitution en ne prenant en compte qu'un certain nombre d'axes et de composantes, ce qui améliore les traitements statistiques postérieurs à l'ACP (le clustering par exemple). Dans ce cas, l'ACP peut être vue comme une méthode de débruitage des données transcriptomiques. C'est pourquoi nous proposons un nouveau point de vue pour présenter l'ACP qui est basé sur un modèle faisant clairement apparaître les aspects de débruitage. Enfin, ce formalisme nous conduira à proposer une version régularisée de l'ACP permettant d'améliorer le débruitage des données transcriptomiques.

1 Visualisation des données transcriptomiques

Premièrement, nous nous intéressons à l'utilisation de l'ACP en tant qu'outil de visualisation des données transcriptomiques.

1.1 Bilan des pratiques

Dans la très grande majorité des cas, l'ACP est réalisée sur le tableau croisant les sujets et les gènes, ainsi les sujets sont considérés comme les individus et les gènes comme les variables. De plus, précisons que la plupart du temps, seul un sous-ensemble des gènes, dits différentiellement exprimés en fonction des conditions expérimentales, est conservé suite à une procédure de tests multiples. L'interprétation est très souvent succincte et restreinte à la représentation des individus. L'unique interprétation consiste à vérifier que les individus se séparent en fonction des conditions expérimentales sur les premiers axes de l'ACP (par exemple Horinouchi *et al.* (2010); Devonshire *et al.* (2010); Kwekel *et al.* (2010); Mirbahai *et al.* (2011)). Dans ce cas, nous perdons l'essence même de l'ACP que sont les relations de dualité. Il est en effet primordial d'interpréter conjointement les représentations des individus et des variables pour conserver la richesse d'interprétation.

Ainsi, nous nous proposons ici de discuter ces règles d'interprétation.

1.2 Adaptation des règles d'interprétation de l'ACP à l'analyse des données transcriptomiques

A la manière de Baccini *et al.* (2005), nous présentons ici le fruit de nos questionnements et réflexions sur les règles d'interprétation de l'ACP appliquées aux données

transcriptomiques. Nous nous attachons particulièrement à prendre en compte la nature des données transcriptomiques. Nous considérons ici l'ACP du tableau sujets×gènes où les sujets sont considérés comme les individus et les gènes comme les variables, même si nous nous réservons le droit de questionner ce choix par la suite.

1.2.1 Centrage, réduction

En ACP, se pose la traditionnelle question du centrage et de la réduction des variables (traditionnellement les gènes). Le centrage est usuel en ACP et est une opération neutre puisqu'il ne modifie ni les distances relatives entre individus, ni les liaisons entre variables. Le centrage est donc un intermédiaire technique qui offre d'intéressantes propriétés pour l'interprétation de l'ACP et qui est systématiquement réalisé.

La réduction des variables est une question plus délicate et dépend intrinsèquement de la nature des mesures. Dans les données transcriptomiques, la mesure reflète une quantité d'ARNm produit pour un gène donné. Or cette quantité dépend du gène. Si nous prenons l'exemple d'un neurotransmetteur tel que la sérotonine, même si l'expression du gène codant pour la sérotonine est très fortement activée, la quantité d'ARNm produite reste faible puisque ce neurotransmetteur n'est produit qu'en petites quantités par définition. Au contraire si nous prenons l'exemple d'une protéine du cytosquelette telle que l'actine qui est omniprésente, une grande quantité d'ARNm correspondant au gène codant pour cette protéine est systématiquement produite. Nous pouvons donc considérer que la quantité d'ARNm n'a pas la même signification d'un gène à l'autre et représente éventuellement des unités de mesure différentes. Cet argument plaide donc en faveur de la réduction des variables.

De plus, comme très souvent dans l'expérimentation, nous apportons des jugements relatifs en fonction des conditions expérimentales. Autrement dit, nous ne nous intéressons pas aux quantités en tant que telles mais aux variations des quantités entre les différentes conditions expérimentales.

1.2.2 Orientation du tableau de données

En ACP, le problème de la dimensionnalité n'apparaît pas lorsque le nombre de variables est beaucoup plus grand que le nombre d'individus comme en régression par exemple. L'ACP peut donc être appliquée à des jeux de données où le nombre de variables est très supérieur au nombre d'individus. Comme nous l'avons établi, l'ACP est usuellement appliquée au tableau sujets×gènes, où les sujets sont considérés comme les individus et les gènes comme les variables. Nous pouvons néanmoins questionner l'orientation du tableau de données.

Dans le cadre classique, un sujet est considéré comme un ensemble de valeurs d'expression que l'on nomme profil d'expression. Les données sont considérées comme centrées et réduites par gène. Dans cette ACP, deux sujets sont d'autant plus proches qu'ils présentent des profils d'expression similaires. Deux gènes sont d'autant plus corrélés qu'ils s'expriment de la même manière chez les sujets.

Nous pouvons également réaliser l'ACP du tableau gènes×sujets car un gène peut aussi être considéré comme un ensemble de valeurs d'expression qui mesurent la réaction à un ensemble de conditions expérimentales. Dans cette transposition du tableau, nous conservons le centrage réduction par gène, donc par ligne (pour les raisons évoquées section 1.2.1). S'ajoute à cela un centrage par sujet, qui est nécessaire en ACP pour interpréter les axes comme des dimensions de variabilité au sein des individus. Ce centrage est neutre comme nous l'avons précédemment fait remarquer (section 1.2.1).

Afin de confronter ces deux points de vue, considérons l'ACP des deux tableaux issus de notre jeu de données transcriptomiques illustratif (présenté chapitre 1 section 3). Si nous comparons la représentation des individus de l'ACP sujets×gènes (figure 1) avec la représentation des variables de l'ACP gènes×sujets (figure 2), nous pouvons clairement remarquer la grande similarité de ces deux représentations. Il en est de même si nous comparons la représentation des variables de l'ACP sujets×gènes avec la représentation des individus de l'ACP gènes×sujets, ce qui est attendu du fait de la dualité. D'ailleurs, le coefficient RV (Escoufier, 1973) entre les deux ensembles de coordonnées est de 0.99. Ainsi, les distances ne sont pas parfaitement équivalentes sur les premiers plans factoriels. Nous analysons bien le même nuage des gènes dans ces deux ACP, dans un cas par rapport à l'origine (ACP sujets×gènes), dans l'autre par rapport à son centre de gravité du fait du centrage (ACP gènes×sujets). Ainsi, si le centre de gravité du nuage est très proche de l'origine, nous obtenons des premiers plans comparables, ce qui est le cas ici : la distance entre l'origine et le centre de gravité est de 0.05 ce qui est petit devant 1, d'où la valeur de 0.99 du coefficient RV. En d'autres termes, s'il existe un fort effet taille, les deux ACP mènent à des règles d'interprétation différentes.

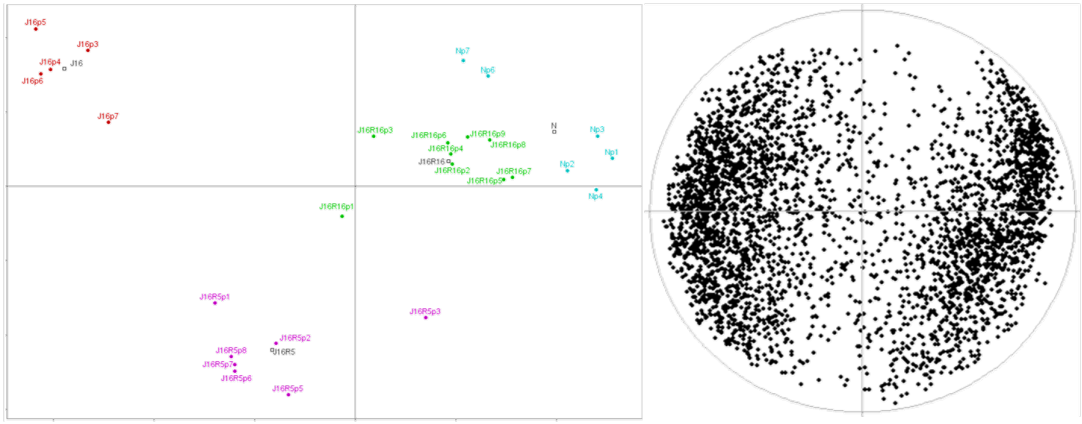


FIGURE 1 – Premier plan factoriel de l'ACP sujets×gènes du jeu de données transcriptomiques illustratif (présenté chapitre 1 section 3) : représentation des individus (sujets, à gauche) et des variables (gènes, à droite).

L'ACP du tableau sujets×gènes fournit des règles d'interprétation aisées. Si pour un gène donné et pour un sujet donné une valeur d'expression est positive, nous dirons que le gène est sur-exprimé chez le sujet, au contraire, si cette valeur est négative, le gène est

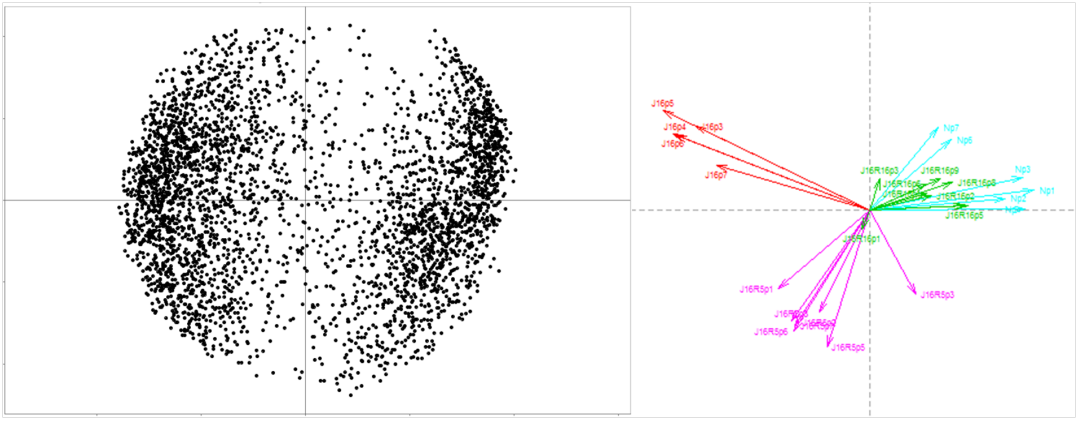


FIGURE 2 – Premier plan factoriel de l'ACP gènes×sujets du jeu de données transcriptomiques illustratif (présenté chapitre 1 section 3) : représentation des individus (gènes, à gauche) et des variables (sujets, à droite).

dit sous-exprimé chez le sujet. Néanmoins, l'ACP du tableau gènes×sujets permet d'introduire facilement de l'information sur les gènes. En ce sens elle peut être un intermédiaire technique intéressant (voir chapitre 3).

1.2.3 Sélection de gènes

L'ACP comme toutes les méthodes d'analyse factorielle présente un intérêt pour les données qui présentent un caractère exhaustif (Benzecri, 1982), ce qui est le cas dans les recueils de données transcriptomiques où l'on peut s'intéresser à l'ensemble des gènes qui s'expriment dans un tissu donné par exemple. Dans notre jeu de données illustratif, nous nous intéressons à l'ensemble des gènes dits d'expression hépatique chez le poulet. Cependant, nous recueillons inévitablement de l'information sans rapport avec les conditions expérimentales. Se pose alors la question de la sélection de gènes.

Une première idée consiste dans un cadre d'analyse factorielle à éliminer les gènes de variance nulle, et à adopter une démarche que nous pourrions qualifier d'exhaustivité recentrée. Nous pouvons également choisir de renforcer la structure du plan factoriel en effectuant une sélection des gènes qui sont différentiellement exprimés en fonction des conditions expérimentales. Cependant, la structuration du plan factoriel en fonction des conditions expérimentales, mécanique à l'issue de la sélection, peut être intéressante à observer dans un cadre exhaustif. Ce faisant, nous orientons la construction du plan factoriel vers la discrimination des conditions expérimentales, auquel cas le fait d'observer des premiers axes factoriels en fonction des conditions expérimentales est attendu. En revanche, si nous conservons tous les gènes dans l'analyse, situer les axes induits par les conditions expérimentales parmi d'autres est une information précieuse. En faisant l'ACP des seuls gènes différentiellement exprimés en fonction des conditions expérimentales, nous nous rapprochons d'une analyse factorielle discriminante qui pousserait à l'extrême l'idée d'extraire des axes discriminant les sujets uniquement en fonction des différentes

conditions expérimentales. Ainsi, nous pouvons voir la sélection comme un compromis entre une analyse factorielle sur l'intégralité des gènes et une analyse factorielle discriminante, dans la mesure où nous privilégions les facteurs expérimentaux, mais sans interdire l'expression d'éventuels facteurs parasites dont la détection et l'évaluation par rapport aux facteurs expérimentaux sont indispensables. En conclusion, il peut être judicieux de réaliser à la fois l'ACP du jeu de données complet et du jeu de données issu d'une sélection des gènes.

Après ces considérations générales sur l'utilisation de l'ACP dans la visualisation des données transcriptomiques, nous proposons un exemple d'interprétation.

1.3 Exemple d'interprétation

Nous proposons de fournir un exemple d'interprétation des sorties de l'ACP sur notre jeu de données illustratif. Nous souhaitons illustrer à travers cet exemple comment à partir de l'interprétation générale des axes de l'ACP, nous pouvons déduire des hypothèses sur l'expérience. L'ACP est appliquée au tableau sujets$\times$gènes, les gènes étant issus d'une sélection par tests multiples et centrés et réduits dans cette analyse (figure 1).

Nous nous attendons à retrouver un premier axe factoriel prépondérant dans lequel les conditions expérimentales s'ordonnent depuis le jeûne jusqu'à l'état nourri. C'est effectivement ce que nous obtenons. Le pourcentage d'inertie de 35% donne l'importance relative du premier axe : il y a d'autres sources de variabilité non négligeables, dues à la non linéarité d'une part, il s'agit des axes 2 et 3, et à la variabilité intra condition expérimentale d'autre part. Dans le détail, les écarts entre les modalités ne sont pas les mêmes sur cet axe. Ils suggèrent, en particulier, qu'au bout de 16h de renutrition, les effets du jeûne sont fortement atténués, mais pas annulés. De plus, la modalité J16R5 semble équidistante des modalités J16 et J16R16, ce qui suggère que la récupération est plus rapide au début de la période de renutrition.

Sur le premier plan factoriel, le nuage des individus présente une structure tripolaire. Les individus nourris (N) et à jeun 16h puis renourris 16h (J16R16) s'opposent aux individus à jeun (J16) sur l'axe 1, tandis que les individus à jeun 16h puis renourris 5h (J16R5) s'opposent aux individus J16 sur l'axe 2. La structure triangulaire du nuage des individus traduit l'existence de gènes activés ou réprimés spécifiquement dans chacun des états nutritionnels. Du fait de la dualité, nous pouvons nous attendre à observer une structure tripolaire des variables, chaque pôle correspondant à l'un des trois pôles d'individus. Or ce n'est pas le cas : la structure du nuage sur le cercle des corrélations est particulière puisque nous observons un vide. Nous déduisons du vide observé que peu de gènes sont exprimés ou inhibés spécifiquement chez les individus J16R5. Nous pouvons penser que chez ces individus, à la fois il reste des $\ll$ traces $\gg$ du jeûne, et le métabolisme basal est réactivé. Ceci expliquerait que de nombreux gènes soient exprimés à la fois chez les individus J16 et J16R5 et chez les individus N et J16R5.

Grâce à cet exemple, nous montrons qu'il est possible de tirer beaucoup d'informations de l'interprétation générale d'une ACP, notamment en utilisant les relations de dualité.

Outre la visualisation de données, l'ACP peut être vue comme une méthode permettant de réduire la dimension des données, soit en ne s'intéressant à l'interprétation que des premières dimensions, soit en obtenant un jeu de données en dimension inférieure à partir des premiers axes et composantes, au moyen de la formule de reconstitution. Réduire ainsi la dimension des données peut être interpréter comme un pré-traitement destiné à débruiter les données (Alter *et al.*, 2000; Raychaudhuri *et al.*, 2000). En effet, nous définissons ainsi le débruitage des données comme un traitement qui permet de séparer un signal (ou des grandes tendances) que nous ne connaissons pas mais qui est l'objet d'intérêt, d'un bruit que nous cherchons à atténuer. Nous nous intéressons donc à cette seconde utilisation de l'ACP, en tant qu'outil de pré-traitement, dans l'analyse des données transcriptomiques.

2 Débruitage des données transcriptomiques

2.1 Bilan des pratiques

Tout d'abord, nous pouvons citer plusieurs types de traitements qui sont précédés d'une ACP dans le but de débruiter les données, tels que le clustering ou encore l'inférence de réseaux de régulation de gènes.

2.1.1 Indicateurs synthétiques

Sachant que les données transcriptomiques sont des données, dites en grande dimension, dans la littérature plusieurs travaux proposent d'utiliser les variables synthétiques obtenues par ACP à la place des variables originales. Dans ce cas, nous nous intéressons aux variables synthétiques et non aux gènes pour interpréter les données et formuler des hypothèses. Cette démarche peut être interprétée comme un débruitage dans le sens où l'on ne s'intéresse qu'aux grandes tendances, qui représentent le signal que nous cherchons à interpréter, tout en éliminant les tendances particulières qui ne sont pas l'objet d'intérêt et qui vont même jusqu'à perturber l'analyse.

Nous pouvons donner l'exemple de Holter *et al.* (2000) qui utilisent l'ACP pour représenter un petit nombre de groupes d'expressions géniques qui varient de la même façon en fonction du temps. Autrement dit chaque composante principale, nommée « characteristic mode », représente un signal temporel d'expression. Étudier les expressions géniques à travers les « characteristic modes » au lieu de les étudier à partir des expressions seules permet de mieux comprendre les phénomènes biologiques de l'expérience.

A partir de ces mêmes données transcriptomiques temporelles, Holter *et al.* (2001) proposent à travers une matrice de traduction temporelle, de prédire le niveau d'expression des gènes dans un temps futur à partir du niveau d'expression au temps initial. Prendre en compte les characteristic modes à la place des signaux bruts améliore la prédiction du niveau d'expression en débruitant les données.

2.1.2 Clustering

Le clustering est fréquemment utilisé soit pour obtenir des clusters de gènes coexprimés, ou encore pour visualiser le tableau de données à travers une double classification ascendante hiérarchique à la fois sur les individus et sur les variables comme dans les heatmaps (Eisen *et al.*, 1998). Il est fréquent d'utiliser une ACP comme pré-traitement à un clustering. Dans ce cas, l'ACP est utilisée pour obtenir un tableau de données reconstituées à partir des seules premières dimensions avant d'appliquer un algorithme de clustering (Alter *et al.*, 2000). Procéder ainsi permet d'améliorer les partitions obtenues qui sont plus cohérentes (Yeung et Ruzzo, 2001). Nous pouvons également citer deux méthodes de clustering de gènes (Hastie *et al.*, 2000; Wall *et al.*, 2001) dans lesquelles une étape d'ACP est inclue dans l'algorithme de clustering, ce qui permet d'obtenir des clusters plus interprétables.

2.1.3 Inférence de réseaux de régulation

Dans un cas d'inférence de réseaux de régulation, une étape intermédiaire à l'inférence de réseaux est souvent une étape d'estimation de la matrice de variance-covariance des gènes. Le fait que le nombre de gènes soit bien plus grand que le nombre d'individus rend l'estimation de cette matrice de variance-covariance très peu stable. Ainsi estimer la matrice de variance-covariance à partir d'une matrice reconstituée en ne conservant que les premières composantes permet d'obtenir une estimation plus stable. Les réseaux estimés à partir des matrices de variance-covariance, que l'on peut qualifier de débruitées, sont de meilleure qualité (Yeung *et al.*, 2002).

2.2 Vers un modèle signal + bruit

L'ACP étant fréquemment utilisée comme un outil de débruitage des données transcriptomiques, cela sous-entend que les données transcriptomiques peuvent être considérées comme un mélange de signal d'intérêt mais que l'on ne connaît pas et de bruit que l'on cherche à éliminer. Tout d'abord, nous nous interrogeons sur la plausibilité d'une telle hypothèse sur les données transcriptomiques.

2.2.1 Nature bruitée des données transcriptomiques

Les données transcriptomiques sont donc usuellement considérées comme bruitées. Précisons que le bruit peut être artificiellement séparé en deux concepts. D'une part, il

existe du bruit lié à une erreur de mesure, ce qui produit une certaine incertitude sur les valeurs du jeu de données. Dans les recueils de données transcriptomiques, plusieurs sources de variabilité technologique peuvent être à l'origine d'une incertitude sur les mesures. Effectivement, à chaque étape du recueil de données, de la variabilité technologique peut être à l'origine d'un bruit technologique (section 1.3.2 chapitre 1). Nous pouvons par exemple citer la préparation des échantillons qui peut varier d'un échantillon à l'autre, ou encore le lavage de la puce plus ou moins stringent. Précisons que les données que nous interprétons en général, ont déjà subi une étape de débruitage à travers un traitement des images de fluorescence brute qui réduit le bruit d'origine technologique. D'autre part, nous pouvons identifier du bruit lié à la variabilité inter-individuelle. En effet cette variabilité est particulièrement inéluctable lorsque l'on traite des données biologiques. Il est donc indispensable de lisser ces variations individuelles afin de focaliser l'interprétation sur les grandes tendances.

Nous avons donc établi que les données transcriptomiques peuvent être considérées comme bruitée et l'ACP est couramment utilisée comme un outil de débruitage. Ces considérations nous ont amenés à nous intéresser à la définition d'un modèle pour l'ACP.

2.2.2 Point de vue modèle sur l'ACP

Classiquement l'ACP d'une matrice de données $\mathbf{X}$ quelconque est présentée à travers un point de vue géométrique et fournit des axes et des composantes. Comme nous l'avons déjà dit, à partir des axes et des composantes et au moyen de la formule de reconstitution, il est possible d'estimer un tableau de données, que nous notons $\hat{\mathbf{X}}$. Sachant cela, nous pouvons également définir l'ACP à travers le critère des moindres carrés qui revient à estimer $\hat{\mathbf{X}}$ en minimisant l'erreur de reconstitution $\|\mathbf{X} - \hat{\mathbf{X}}\|^2$. Si nous considérons une estimation de la matrice de données $\hat{\mathbf{X}}$ à partir des S premières composantes uniquement, l'ACP permet une approximation de $\mathbf{X}$ en rang inférieur S.

Nous pouvons également adopter un point de vue modèle sur l'ACP. Un modèle classique pour l'ACP est le modèle à effets fixes introduit par Caussinus (1986) qui est un modèle de type signal + bruit et qui se définit ainsi :

$$\begin{aligned} \mathbf{X} &= \tilde{\mathbf{X}} + \varepsilon \\ x_{ij} &= \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{1}$$

Dans ce modèle une matrice de données $\mathbf{X}$ (supposée centrée) peut se décomposer en un signal $\tilde{\mathbf{X}}$ auquel s'ajoute un bruit gaussien ε. Sous sa forme indicée, nous notons d_s la $s^{\text{ème}}$ valeur propre de $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ (n fois la « vraie », celle du signal, matrice de variance-covariance), $\mathbf{r_s} = \{r_{1s}, ..., r_{js}, ..., r_{ps}\}$ le vecteur propre associé et $\mathbf{q_s} = \{q_{1s}, ..., q_{is}, ..., q_{ns}\}$ le $s^{\text{ème}}$ vecteur propre de $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ (n fois la vraie matrice de produits scalaires). Précisons que l'estimateur du maximum de vraisemblance de ce modèle correspond bien à l'estimateur des moindres carrés précédemment défini. Considérer l'ACP sous forme de modèle (1) fait apparaître la propriété de débruitage de l'ACP. Ce nouveau point de vue nous entraîne

à nous interroger sur cette propriété de débruitage. En effet, alors que l'ACP consiste à minimiser le critère des moindres carrés du type $\|\mathbf{X} - \hat{\mathbf{X}}\|^2$ qui revient à minimiser l'écart entre les données et l'estimateur, le fait d'expliciter le modèle nous incite à considérer un critère du type $\|\tilde{\mathbf{X}} - \hat{\mathbf{X}}\|^2$ qui revient à minimiser l'écart entre le signal et l'estimateur. Or ce critère est proche de l'erreur quadratique moyenne (2) :

$$\text{EQM} = \mathbb{E}\left(\sum_{i,j}(\hat{x}_{ij} - \tilde{x}_{ij})^2\right) \tag{2}$$

Cependant, il est établi en régression par exemple que les estimateurs du maximum de vraisemblance ne sont pas les meilleurs pour minimiser l'erreur quadratique moyenne, mais des estimateurs régularisés permettent de réduire cette erreur. Ainsi nous avons mis au point une stratégie de régularisation de l'ACP. L'ACP régularisée est proposée dans l'article suivant :

Verbanck, M., Josse, J., & Husson, F. (2013). Regularised PCA to visualise and denoise data. *Statistics and Computing* (soumis)

Regularised PCA to denoise and visualise data (Verbanck *et al.*, 2013a)

Résumé : L'ACP est une méthode classique qui peut être utilisée comme méthode de réduction de la dimension afin de débruiter des données. Si nous considérons le modèle à effets fixes de l'ACP qui décompose une matrice de données en un signal plus du bruit, nous pouvons montrer que l'ACP ne fournit pas la meilleure reconstitution du signal en termes d'erreur quadratique moyenne. Suivant le même principe qu'en régression, nous proposons une version régularisée de l'ACP qui revient à sélectionner un certain nombre de dimensions et à régulariser les valeurs singulières correspondantes. Dans la formule de reconstitution des données, chaque valeur singulière est ainsi multipliée par un terme qui peut être vu comme le ratio de la variance du signal sur la variance totale pour la dimension correspondante. Le terme de régularisation est dérivé analytiquement en utilisant des résultats asymptotiques et peut également être justifié comme un traitement bayésien du modèle à effets fixes. L'ACP régularisée fournit des résultats encourageants en termes de reconstitution du vrai signal et des sorties graphiques en comparaison avec l'ACP classique et une méthode de seuillage doux. La distinction entre ACP et ACP régularisée devient encore plus importante dans le cas de données très bruitées.

Noname manuscript No.
(will be inserted by the editor)

Regularised PCA to denoise and visualise data

Marie Verbanck · Julie Josse · François Husson

Received: date / Accepted: date

Abstract Principal component analysis (PCA) is a well-established dimensionality reduction method commonly used to denoise and visualise data. A classical PCA model is the fixed effect model in which data are generated as a fixed structure of low rank corrupted by noise. Under this model, PCA does not provide the best recovery of the underlying signal in terms of mean squared error. Following the same principle as in ridge regression, we suggest a regularised version of PCA that essentially selects a certain number of dimensions and shrinks the corresponding singular values. Each singular value is multiplied by a term which can be seen as the ratio of the signal variance over the total variance of the associated dimension. The regularised term is analytically derived using asymptotic results and can also be justified from a Bayesian treatment of the model. Regularised PCA provides promising results in terms of the recovery of the true signal and the graphical outputs in comparison with classical PCA and with a soft thresholding estimation strategy. The distinction between PCA and regularised PCA becomes especially important in the case of very noisy data.

Keywords principal component analysis · shrinkage · regularised PCA · fixed effect model · denoising · visualisation

M. Verbanck
Applied mathematics department, Agrocampus Ouest
Tel.: +332-23-48-54-91
E-mail: marie.verbanck@agrocampus-ouest.fr

J. Josse
Applied mathematics department, Agrocampus Ouest
Tel.: +332-23-48-58-74
E-mail: julie.josse@agrocampus-ouest.fr

F. Husson
Applied mathematics department, Agrocampus Ouest
Tel.: +332-23-48-58-86
E-mail: francois.husson@agrocampus-ouest.fr

1 Introduction

In many applications (Mazumder et al, 2010; Candès et al, 2012), we can consider that data are generated as a structure having a low rank representation corrupted by noise. Thus, the associated model for any data matrix $\mathbf{X}$ (assumed without loss of generality to be centered) composed of n individuals and p variables can be written as:

$$\mathbf{X}_{n\times p} = \tilde{\mathbf{X}}_{n\times p} + \varepsilon_{n\times p} \quad (1)$$

$$x_{ij} = \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

where d_s is the s^{th} eigenvalue of the matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ (n times the true covariance matrix), $\mathbf{r}_s = \{r_{1s}, ..., r_{js}, ..., r_{ps}\}$ is the associated eigenvector and $\mathbf{q}_s = \{q_{1s}, ..., q_{is}, ..., q_{ns}\}$ is the s^{th} eigenvector of the matrix $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ (n times the true inner-product matrix). Such a model is also known as the fixed effect model (Caussinus, 1986) in principal component analysis (PCA).

PCA is a well-established dimensionality reduction method. It allows the data $\mathbf{X}$ to be described using a small number (S) of uncorrelated variables (the principal components) while retaining as much information as possible. PCA is often used as an exploratory method to summarise and visualise data. PCA is also often considered as a way of separating the signal from the noise where the first S principal components are taken as the signal while the remaining ones as the noise. Therefore, PCA can be used as a denoising method to analyse images for instance or to preprocess data before applying other methods such as clustering. Indeed, clustering is expected to be more stable when applied to noise-free data sets.

PCA provides a subspace which best represents the data, that is, which minimises the distances between individuals and their projection on the subspace. Formally, this corresponds to finding a matrix $\hat{\mathbf{X}}_{n\times p}$, of low rank S, which minimises $\|\mathbf{X} - \hat{\mathbf{X}}\|^2$ with $\| \bullet \|$ the Frobenius norm. The solution is given by the singular value decomposition (SVD) of $\mathbf{X}$:

$$\hat{x}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} \qquad (2)$$

where λ_s is the s^{th} eigenvalue of $\mathbf{X}'\mathbf{X}$, $\mathbf{u}_s = \{u_{1s}, ..., u_{is}, ..., u_{ns}\}$ the s^{th} left singular vector and $\mathbf{v}_s = \{v_{1s}, ..., v_{js}, ..., v_{ps}\}$ the s^{th} right singular vector. This least squares estimator corresponds to the maximum likelihood solution of model (1).

It is established, for instance in regression, that the maximum likelihood estimators are not necessarily the best for minimising mean squared error (MSE). However, shrinkage estimators, although biased, have smaller variance which may reduce the MSE. We follow this approach and propose a regularised version of PCA in order to get a better estimate of the underlying structure $\tilde{\mathbf{X}}$. In addition, this approach allows graphical representations which are as close as possible to the representations that would be obtained from the signal only. As we will show later, our approach essentially shrinks the first S singular values with a different amount of shrinkage for each singular value. The shrinkage terms will be analytically derived.

In the literature, a popular strategy to recover a low rank signal from noisy data is to use a soft thresholding strategy. More precisely, each singular value is thresholded with a constant amount of shrinkage usually found by cross-validation. However, recently, Candès et al (2012) suggested determining the threshold level without resorting to a computational method by minimising an estimate of the risk, namely a Stein's unbiased risk estimate (SURE). We will compare our approach to this SURE method.

In this paper, we derive the shrinkage terms by minimising the mean squared error and define regularised PCA (rPCA) in Section 2. We also show that rPCA can be derived from a Bayesian treatment of the fixed effect model (1). Section 3 shows the efficiency of regularisation through a simulation study in which rPCA is compared to classical PCA and the SURE method. The performance of rPCA is illustrated through the recovery of the signal and the graphical outputs (individual and variable representations). Finally, rPCA is performed on a real microarray data set and on images in Section 4.

2 Regularised PCA

2.1 MSE point of view

2.1.1 Minimising the MSE

PCA provides an estimator $\hat{\mathbf{X}}$ which is as close as possible to $\mathbf{X}$ in the least squares sense.

However, assuming model (1), the objective is to get an estimator as close as possible to the unknown signal $\tilde{\mathbf{X}}$. To achieve such a goal, the same principle as in ridge regression is followed. We look for a shrinkage version of the maximum likelihood estimator which is as close as possible to the true structure. More precisely, we look for shrinkage terms $\mathbf{\Phi} = (\phi_s)_{s=1,\ldots,\min(n-1,p)}$ that minimise:

$$\text{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{\min(n-1,p)} \phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}^{(s)}\right)^2\right)$$
$$\text{with } \hat{x}_{ij}^{(s)} = \sqrt{\lambda_s} u_{is} v_{js};\ \tilde{x}_{ij}^{(s)} = \sqrt{d_s} q_{is} r_{js}$$

First, we separate the terms of the MSE corresponding to the first S dimensions from the remaining ones:

$$\text{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}^{(s)}\right)^2 + \left(\sum_{s=S+1}^{\min(n-1,p)} \phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}^{(s)}\right)^2\right)$$

Then, according to equation (1), for all $s \geq S+1$, $\tilde{x}_{ij}^{(s)} = 0$. Therefore, the MSE is minimised for $\phi_{S+1} = \ldots = \phi_{\min(n-1,p)} = 0$. Thus, the MSE can be written as:

$$\text{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}^{(s)}\right)^2\right)$$

Using the orthogonality constraints, for all $s \neq s'$, $\sum_i u_{is}u_{is'} = \sum_j v_{js}v_{js'} = 0$, the MSE can be simplified as follows:

$$\text{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s^2 \lambda_s u_{is}^2 v_{js}^2 - 2\tilde{x}_{ij}\sum_{s=1}^{S} \phi_s \sqrt{\lambda_s} u_{is} v_{js} + (\tilde{x}_{ij})^2\right)\right) \quad (3)$$

Finally, equation (3) is differentiated with respect to ϕ_s to get:

$$\phi_s = \frac{\sum_{i,j} \mathbb{E}\left(\hat{x}_{ij}^{(s)}\right)\tilde{x}_{ij}}{\sum_{i,j} \mathbb{E}\left(\hat{x}_{ij}^{(s)2}\right)} = \frac{\sum_{i,j} \mathbb{E}\left(\hat{x}_{ij}^{(s)}\right)\tilde{x}_{ij}}{\sum_{i,j}\left(\mathbb{V}\left(\hat{x}_{ij}^{(s)}\right) + \left(\mathbb{E}\left(\hat{x}_{ij}^{(s)}\right)\right)^2\right)}$$

Then, to simplify this quantity, we adapt results coming from the setup of analysis of variance with two factors to the PCA framework. More precisely, we use the results of Denis and Pázman (1999) and Denis and Gower (1996) who studied nonlinear regression models with constraints and focused on bilinear models called biadditive models. Such models are defined as follow:

$$y_{ij} = \mu + \alpha_i + \beta_j + \sum_{s=1}^{S} \gamma_{is}\delta_{js} + \varepsilon_{ij} \quad (4)$$
$$\text{with } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

where y_{ij} is the response for the category i of the first factor and the category j of the second factor, μ is the grand mean, $(\alpha_i)_{i=1,\ldots,I}$ and $(\beta_j)_{j=1,\ldots,J}$ correspond to the main effect parameters and $\left(\sum_{s=1}^{S} \gamma_{is}\delta_{js}\right)_{i=1,\ldots,I;j=1,\ldots,J}$ model the interaction. The least squares estimates of the multiplicative terms are given by the singular value decomposition of the residual matrix of the model without interaction. From a computational point of view, this model is similar to the PCA one, the main difference being that the linear part only includes the grand mean and column main effect in PCA. Using the Jacobians and the Hessians of the response defined by Denis and Gower (1994) and recently in Papadopoulo and Lourakis (2000), Denis and Gower (1996) derived the asymptotic bias of the response of model (4) and showed that the response estimator is approximately unbiased. Transposed to the PCA framework, it leads to conclude that the PCA estimator is asymptotically unbiased $\mathbb{E}\left(\hat{x}_{ij}\right) = \tilde{x}_{ij}$ and for each dimension s, $\mathbb{E}\left(\hat{x}_{ij}^{(s)}\right) = \tilde{x}_{ij}^{(s)}$. In

addition, the variance of $\hat{x}_{ij}$ can be approximated by the noise variance. Therefore, we estimate $\mathbb{V}\left(\hat{x}_{ij}^{(s)}\right)$ by the average variance, that is $\mathbb{V}\left(\hat{x}_{ij}^{(s)}\right) = \frac{1}{\min(n-1;p)}\sigma^2$.

Consequently ϕ_s can be approximated by:

$$\phi_s = \frac{\sum_{i,j} \tilde{x}_{ij}^{(s)} \tilde{x}_{ij}}{\sum_{i,j}\left(\frac{1}{\min(n-1;p)}\sigma^2 + (\tilde{x}_{ij}^{(s)})^2\right)}$$

Since for all $s \neq s'$, the dimensions s and s' of $\tilde{\mathbf{X}}$ are orthogonal, thus ϕ_s can be written as:

$$\phi_s = \frac{\sum_{i,j} \tilde{x}_{ij}^{(s)2}}{\sum_{i,j}\left(\frac{1}{\min(n-1;p)}\sigma^2 + (\tilde{x}_{ij}^{(s)})^2\right)}$$

Based on equation (1), the quantity $\sum_{i,j}(\tilde{x}_{ij}^{(s)})^2$ is equal to d_s the variance of the s^{th} dimension of the signal. ϕ_s is then equal to:

$$\phi_s = \begin{cases} \dfrac{d_s}{\frac{np}{min\{p,n-1\}}\sigma^2 + d_s} \forall s = 1, ..., S \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \quad (5)$$

The form of the shrinkage term is appealing since it corresponds to the ratio of the variance of the signal over the total variance (signal plus noise) for the s^{th} dimension.

Remark: Models such as model (4) *are also known as additive main effects and multiplicative interaction (AMMI) models. They are often used to analyse genotype-environment data in plant breading framework. Considering a random version of such models, Cornelius and Crossa (1999) developed a regularisation term which is similar to ours. It allows improved prediction of the yield obtained by genotypes in environments*

2.1.2 Definition of regularised PCA

The shrinkage terms (5) depend on unknown quantities. We estimate them by plug-in. The total variance of the s^{th} dimension is estimated by the variance of $\mathbf{X}$ for the dimension s, *i.e.* by its associated eigenvalue λ_s. The signal variance of the s^{th} dimension is estimated by the estimated total variance of the s^{th} dimension minus an estimate of the noise variance of the s^{th} dimension. Consequently, ϕ_s is estimated by $\hat{\phi}_s = \frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}$. Regularised PCA (rPCA) is thus defined by multiplying the maximum likelihood solution by the shrinkage terms which leads to:

$$\begin{aligned} \hat{x}_{ij}^{\text{rPCA}} &= \sum_{s=1}^{S}\left(\frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}\right)\sqrt{\lambda_s}u_{is}v_{js} \\ &= \sum_{s=1}^{S}\left(\sqrt{\lambda_s} - \frac{\frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)u_{is}v_{js} \quad (6) \end{aligned}$$

Using matrix notations, with $\mathbf{U}$ being the matrix of the first S left singular vectors of $\mathbf{X}$, $\mathbf{V}$ being the matrix of the first S right singular vectors of $\mathbf{X}$ and $\mathbf{\Lambda}$ being the diagonal matrix with the associated eigenvalues, the fitted matrix by rPCA is:

$$\hat{\mathbf{X}}^{\text{rPCA}} = \mathbf{U}\hat{\mathbf{\Phi}}\mathbf{\Lambda}^{1/2}\mathbf{V}' \quad (7)$$

rPCA essentially shrinks the first S singular values. It can be interpreted as a compromise between hard and soft thresholding. Hard thresholding consists in selecting a certain number of dimensions S which corresponds to classical PCA (equation 2) whereas soft thresholding consists in thresholding all singular values with the same amount of shrinkage (and without prespecifying the number of dimensions). In rPCA, the s^{th} singular value is less shrunk than the $(s+1)^{th}$ one. This can be interpreted as granting a greater weight to the first dimensions. This behaviour seems desirable. Indeed, the first dimensions can be considered as more stable and trustworthy than the last ones. The regularisation procedure relies more heavily on the less variable dimensions. When $\hat{\sigma}^2$ is small, $\hat{\phi}_s$ is close to 1 and rPCA reduces to standard PCA. When $\hat{\sigma}^2$ is high, $\hat{\phi}_s$ is close to 0 and the values of $\hat{\mathbf{X}}^{\text{rPCA}}$ are close to 0 which corresponds to the average of the variables (in the centered case). From a geometrical point of view, rPCA leads to bring the individuals closer to the centre of gravity.

The regularisation procedure requires estimation of the residual variance σ^2. As the maximum likelihood estimator is biased, another estimator corresponds to the ratio of the residual sum of squares divided by the number of observations minus the number of independent parameters. The latter are equal to $p+\Big((nS-S)-\frac{S(S+1)}{2}\Big)+\Big(pS-\frac{S(S+1)}{2}-S\Big)$, *i.e.* p parameters for the centering, $\Big((nS-S)-\frac{S(S+1)}{2}\Big)$ for the centered and orthonormal left singular vectors and $\Big(pS-\frac{S(S+1)}{2}-S\Big)$ for the orthonormal right singular vectors. This number of parameters can also be calculated as the trace of the projection matrix involved in PCA (Candès and Tao, 2009; Josse and Husson, 2011). Therefore, the residual variance is estimated as:

$$\hat{\sigma}^2 = \frac{\|\mathbf{X}-\hat{\mathbf{X}}\|^2}{np-p-nS-pS+S^2+S} = \frac{\sum_{s=S+1}^{\min(n-1;p)}\lambda_s}{np-p-nS-pS+S^2+S} \quad (8)$$

Contrary to many methods, this classical estimator, namely the residual sum of squares divided by the number of observations minus the number of independent parameters, is still biased. This can be explained by the non-linear form of the model or by the fact that the projection matrix (Josse and Husson, 2011) depends on the data.

2.2 Bayesian points of view

Regularised PCA has been presented and defined via the minimisation of the MSE in section 2.1. However, it is possible to define the method without any reference to MSE, instead using Bayesian considerations. It is well known, in linear regression for instance, that there is equivalence between ridge regression and a Bayesian treatment of the regression model. More precisely, the maximum a posteriori of the regression parameters assuming a Gaussian prior for these parameters corresponds to the ridge estimators (Hastie et al, 2009, p. 64). Following the same rationale, we suggest in this section a Bayesian treatment of the fixed effect model (1).

First, several comments can be made on this model. It is called a "fixed effect" model since the structure is considered fixed. Individuals have different expectations and randomness is only due to the error term. This model is most justified in situations where PCA is performed on data in which the individuals themselves are of interest and are not a random sample drawn from a population of individuals. Such situations frequently arise in practice. For instance, in sensory analysis, individuals can be products, such as chocolates, and variables can be sensory descriptors, such as bitterness, sweetness, etc. The aim is to study these specific products and not others (they are not interchangeable). It thus makes sense to estimate the individual parameters ($\mathbf{q}_s$) and to study the graphical representation of the individuals as well as the representation of the variables. In addition, let us point out that the inferential framework associated with this model is not usual. Indeed the number of parameters increases when the number of individuals increases. Consequently, in this model, asymptotic results are obtained by considering that the noise variance tend to 0.

To suggest a Bayesian treatment of the fixed effect model, we first recall the principle of probabilistic PCA (Roweis, 1998; Tipping and Bishop, 1999) which will be interpreted as a Bayesian treatment of this model.

2.2.1 Probabilistic PCA model

The probabilistic PCA (pPCA) model is a particular case of a factor analysis model (Bartholomew, 1987) with an isotropic noise. The idea behind these models is to summarise the relationships between variables using a small number of latent variables. More precisely, denoting $\mathbf{x}_i$ a row of the matrix $\mathbf{X}$, the pPCA model is written as follows:

$$\mathbf{x}_i = \mathbf{B}_{p\times S}\mathbf{z}_i + \varepsilon_i$$
$$\mathbf{z}_i \sim \mathcal{N}(0, \mathbb{I}_S),\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2\mathbb{I}_p)$$

with $\mathbf{B}_{p\times S}$ being the matrix of unknown coefficients, $\mathbf{z}_i$ being the latent variables and $\mathbb{I}_S$ and $\mathbb{I}_p$ being the identity matrices of size S and p. This model induces a Gaussian distribution on the individuals (which are independent and identically distributed) with a specific structure of variance-covariance:

$$\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Sigma}) \text{ with } \mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \sigma^2\mathbb{I}_p$$

There is an explicit solution for the maximum likelihood estimators:

$$\hat{\mathbf{B}} = \mathbf{V}(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S)^{\frac{1}{2}}\mathbf{R} \tag{9}$$

with $\mathbf{V}$ and $\mathbf{\Lambda}$ defined as in equation (7), that is, as the matrix of the first S left singular vectors of $\mathbf{X}$ and as the diagonal matrix of the eigenvalues, $\mathbf{R}_{S\times S}$ a rotation matrix (usually equal to $\mathbb{I}_S$) and σ^2 estimated as the mean of the last eigenvalues.

In contrast to the fixed effect model (1), the pPCA model can be seen as a random effect model since the structure is random because of the Gaussian distribution on the latent variables. Consequently, this model seems more appropriate when PCA is performed on sample data such as survey data. In such cases, the individuals are not themselves of interest but only considered for the information they provide on the links between variables. Consequently, in such studies, at first, it does not make sense to consider "estimates" of the "individual parameters" since no parameter is associated with the individuals, only random variables ($\mathbf{z}_i$). However, estimators of the "individual parameters" are usually calculated as the expectation of the latent variables given the observed variables $\mathbb{E}(\mathbf{z}_i|\mathbf{x}_i)$. The calculation is detailed in Tipping and Bishop (1999) and results in:

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}} + \sigma^2\mathbb{I}_S)^{-1} \tag{10}$$

We can note that such estimators are often called BLUP estimators (Robinson, 1991) in the framework of mixed effect models where it is also customary to give estimates of the random effects.

Thus, using the maximum likelihood estimator of $\mathbf{B}$ (equation 9) and equation (10), it is possible to build a fitted matrix as:

$$\begin{aligned}\hat{\mathbf{X}}^{\text{pPCA}} &= \hat{\mathbf{Z}}\hat{\mathbf{B}}' = \mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}} + \sigma^2\mathbb{I}_S)^{-1}\hat{\mathbf{B}}' \\ &= \mathbf{X}\mathbf{V}(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S)^{\frac{1}{2}}\mathbf{\Lambda}^{-1}(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S)^{\frac{1}{2}}\mathbf{V}' \\ &= \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S)\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}'\end{aligned}$$

since $\mathbf{X}\ \mathbf{V} = \mathbf{\Lambda}^{1/2}\mathbf{U}$ (given by the SVD of $\mathbf{X}$)

Therefore, considering the pPCA model leads to a fitted matrix of the same form as $\hat{\mathbf{X}}^{\text{rPCA}}$ defined in equation (7) with the same shrunk singular values $\left(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S\right)\mathbf{\Lambda}^{-1/2}$. However, the main difference between the two approaches is that the pPCA model considers individuals as random, whereas they are fixed in model (1) used to define rPCA. Nevertheless, from a conceptual point of view, the random effect model can be considered as a Bayesian treatment of the fixed effect model with a prior distribution on the left singular vectors. Thus, we can consider the pPCA model as the fixed effect model on which we assume a distribution on $\mathbf{z}_i$, considered as the "individual parameters". It is a way to define constraints on the individuals.

Remark: Even if a maximum likelihood solution is available (equation 9) in pPCA, it is possible to use an EM algorithm (Rubin and Thayer, 1982) to estimate the parameters. The two steps correspond to the following two multiple ridge regressions:

Step E: $\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}} + \hat{\sigma}^2\mathbb{I}_S)^{-1}$

Step M: $\hat{\mathbf{B}} = \mathbf{X}'\hat{\mathbf{Z}}(\hat{\mathbf{Z}}'\hat{\mathbf{Z}} + \hat{\sigma}^2\mathbf{\Lambda}^{-1})^{-1}$

Thus, the link between pPCA and the regularised version of PCA is also apparent in these equations. That is, introducing two ridge terms in the two linear multiple regressions which lead to the usual PCA solution (the EM algorithm associated with model (1) *in PCA is also known as the alternative least squares algorithm):*

Step E: $\mathbf{U} = \mathbf{X}\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}$

Step M: $\mathbf{V} = \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}$

2.2.2 An empirical Bayesian approach

Another Bayesian interpretation of the regularized PCA can be given considering directly

an empirical Bayesian treatment of the fixed effect model with a prior distribution on each cell of the data matrix per dimension: $\tilde{x}_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$. From model (1), this implies that $x_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2 + \frac{1}{\min(n-1;p)}\sigma^2)$. The posterior distribution is obtained by combining the likelihood and the priors:

$$
\begin{aligned}
p\left(\tilde{x}_{ij}^{(s)}|x_{ij}^{(s)}\right) &= \frac{p\left(x_{ij}^{(s)}|\tilde{x}_{ij}^{(s)}\right) p\left(\tilde{x}_{ij}^{(s)}\right)}{p\left(x_{ij}^{(s)}\right)} \\
&= \frac{\frac{1}{\sqrt{2\pi\frac{1}{\min(n-1;p)}\sigma^2}} \exp\left[-\frac{\left(x_{ij}^{(s)}-\tilde{x}_{ij}^{(s)}\right)^2}{2\frac{1}{\min(n-1;p)}\sigma^2}\right] \times \frac{1}{\sqrt{2\pi\tau_s^2}} \exp\left[-\frac{\left(\tilde{x}_{ij}^{(s)}\right)^2}{2\tau_s^2}\right]}{\frac{1}{\sqrt{2\pi(\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2)}} \exp\left[-\frac{\left(x_{ij}^{(s)}\right)^2}{2(\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2)}\right]} \\
&= \frac{1}{\sqrt{2\pi\frac{\frac{1}{\min(n-1;p)}\sigma^2\tau_s^2}{\frac{1}{\min(n-1;p)}\sigma^2+\tau_s^2}}} \exp\left[-\frac{\left(\tilde{x}_{ij}^{(s)} - \frac{\tau_s^2}{\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2}\, x_{ij}^{(s)}\right)^2}{2\frac{\frac{1}{\min(n-1;p)}\sigma^2\tau_s^2}{\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2}}\right]
\end{aligned}
$$

The expectation of the posterior distribution is:

$$\mathbb{E}\left(\tilde{x}_{ij}^{(s)}|x_{ij}^{(s)}\right) = \Phi_s x_{ij}^{(s)}$$

$$\text{with } \Phi_s = \frac{\tau_s^2}{\tau_s^2 + \frac{1}{\min(n-1;p)}\sigma^2}$$

This expectation depends on unknown quantities. They are estimated by maximising the likelihood of $\left(x_{ij}^{(s)}\right)_{i=1,\ldots,n;j=1,\ldots,p}$ as a function of τ_s^2 to obtain:

$$\hat{\tau_s}^2 = \left(\frac{1}{np}\lambda_s - \frac{1}{\min(n-1;p)}\hat{\sigma}^2\right)$$

Consequently the shrinkage term is estimated as $\hat{\Phi}_s = \frac{\left(\frac{1}{np}\lambda_s - \frac{1}{\min(n-1;p)}\hat{\sigma}^2\right)}{\frac{1}{np}\lambda_s} = \frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}$ and also corresponds to the regularisation term (6) defined in Section 2.1.1.

Thus, regularised PCA can be seen as a Bayesian treatment of the fixed effect model with a prior on each dimension. The variance of the prior is specific to each dimension s and is estimated as the signal variance of the dimension in question ($\lambda_s - \frac{1}{\min(n-1;p)}\hat{\sigma}^2$).

Remark: Hoff (2007) also proposed a Bayesian treatment of SVD-related models with a primary goal of estimating the number of underlying dimensions. Roughly, his proposition consists in putting prior distributions on $\mathbf{U}$, $\mathbf{\Lambda}$, *and* $\mathbf{V}$. *More precisely, he uses von Mises uniform (Hoff, 2009) prior for orthonormal matrices (on the Steifeld manifold Chikuse (2003)) for* $\mathbf{U}$ *and* $\mathbf{V}$ *and normal priors for the singular values, forming a prior distribution for the structure* $\tilde{\mathbf{X}}$. *Then he builds a Gibbs sampler to get draws from the posterior distributions. The posterior expectation of* $\tilde{\mathbf{X}}$ *can be used as a punctual estimate. It can also be seen as a regularised version of the maximum likelihood estimate. However, contrary to the previously described approach, there is no closed form expression for the regularisation.*

2.3 Bias-variance trade-off

The rationale behind rPCA can be illustrated on graphical representations. Usually, different types of graphical representations are associated with PCA (Greenacre, 2010) depending on whether the left and right singular vectors are represented as normed to 1 or to their associated singular value. In our practice (Husson et al, 2010), we represent the individual

coordinates by $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$ and the variable coordinates by $\mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}$. Therefore, the global shape of the individual cloud represents the variance. Similarly, in the variable representation, the cosine of the angle between two variables can be interpreted as the covariance. Since rPCA ultimately modifies the singular values, it will affect both the representation of the individuals and of the variables. We focus here on the individuals representation.

Data are generated according to model (1) with an underlying signal $\tilde{\mathbf{X}}_{5\times 15}$ composed of 5 individuals and 15 variables in two dimensions. Then, 300 matrices are generated with the same underlying structure: $\mathbf{X}^{sim} = \tilde{\mathbf{X}}_{5\times 15} + \varepsilon^{sim}$ with $sim = 1, ..., 300$. On each data matrix, PCA and rPCA are performed. In figure 1, the configurations of the 5 individuals obtained after each PCA appear on the left, whereas the configurations obtained after each rPCA appear on the right. The average configurations over the 300 simulations are represented by triangles and the true individual configuration obtained from $\tilde{\mathbf{X}}$ is represented by large dots. Representing several sets of coordinates from different PCAs can suffer from translation, reflection, dilatation or rotation ambiguities. Thus, all configurations are superimposed using Procustes rotations (Gower and Dijksterhuis, 2004) by taking as the reference the true individuals configuration.

Compared to PCA, rPCA provides a more biased representation because the coordinates of the average points (triangles) are systematically inferior to the coordinates of the true points (large dots). This is expected because the regularisation term shrinks the individual coordinates towards the origin. In addition, as it is clear for individual number 4 (dark blue), the representation is less variable. Figure 1 thus gives a rough idea of the bias-variance trade-off. Note that even the PCA representation is biased, but this is also expected since $\mathbb{E}(\hat{\mathbf{X}}) = \tilde{\mathbf{X}}$ only asymptotically as detailed in section 2.1.1.

3 Simulation study

To assess rPCA, a simulation study was conducted and rPCA is compared to classical PCA as well as to the SURE method proposed by Candès et al (2012). As explained in the introduction, the SURE method relies on a soft thresholding strategy:

$$\hat{x}_{ij}^{\text{SURE}} = \sum_{s=1}^{\min(n,p)} \left(\sqrt{\lambda_s} - \lambda\right)_+ u_{is} v_{js},$$

The threshold parameter λ is automatically selected by minimising Stein's unbiased risk estimate (SURE). As a tuning parameter, the SURE method does not require the number of underlying dimensions of the signal, but it does require estimation of the noise variance σ^2 to determine λ.

3.1 Recovery of the signal

Data are simulated according to model (1). The structure is simulated by varying several parameters:

- the number of individuals n and the number of variables p based on 3 different combinations: ($n = 100$ and $p = 20$; $n = 50$ and $p = 50$; $n = 20$ and $p = 100$)
- the number of underlying dimensions S (2; 4)
- the ratio of the first eigenvalue to the second eigenvalue (d_1/d_2) of $\tilde{\mathbf{X}}$ (4; 1). When the number of underlying dimensions is higher than 2, the subsequent eigenvalues are roughly of the same order of magnitude.

More precisely, $\tilde{\mathbf{X}}$ is simulated as follows:

1. A SVD is performed on a $n \times S$ matrix generated from a standard multivariate normal distribution. The left singular vectors provide S empirically orthonormal vectors.
2. Each vector $s = 1, ..., S$ is replicated to obtain the p variables. The number of times that each vector s is replicated depends on the ratio between the eigenvalues (d_1/d_2). For instance, if $p = 50$, $S = 2$, $(d_1/d_2) = 4$,

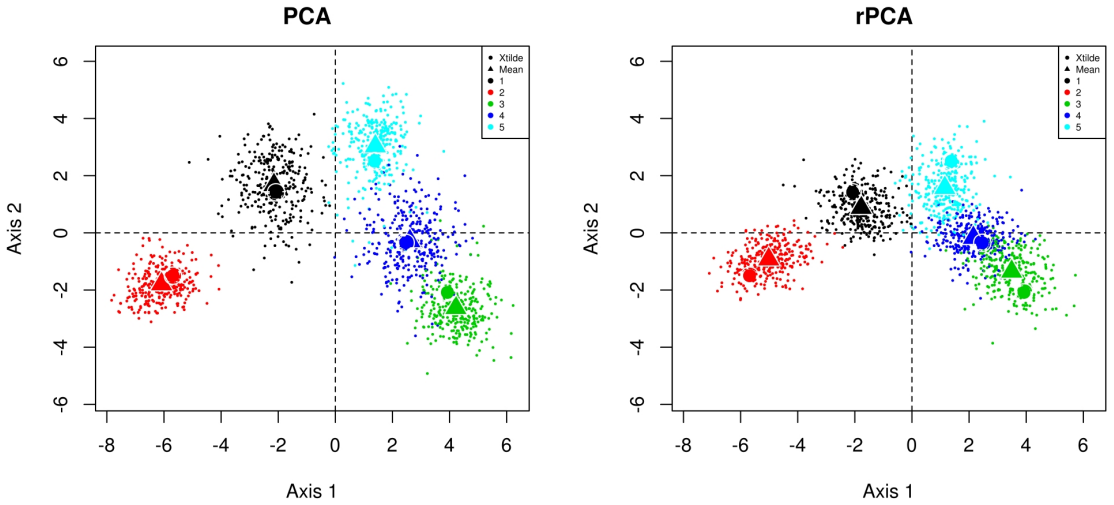


Fig. 1: Superimposition of several configurations of individual coordinates using Procustes rotations towards the true individual configuration of $\tilde{\mathbf{X}}_{5\times 15}$ (large dots). Configurations of the PCA (left) and the rPCA (right) of each $\mathbf{X}^{sim} = \tilde{\mathbf{X}} + \varepsilon^{sim}$, with $sim = 1, ..., 300$ are represented with small dots. The average configuration over the 300 configurations is represented by triangles.

the first vector is replicated 40 times and the second vector is replicated 10 times.

Then, to generate the matrix $\mathbf{X}$, a Gaussian isotropic noise is added to the structure. Different levels of variance σ^2 are considered to obtain three signal-to-noise ratios (Mazumder et al, 2010) equal to 4, 1 and 0.8. A high signal-to-noise ratio (SNR) implies that the variables of $\mathbf{X}$ are very correlated, whereas a low SNR implies that the data are very noisy. For each combination of the parameters, 500 data sets are generated.

To assess the recovery of the signal, the MSE is calculated between the fitted matrix $\hat{\mathbf{X}}$ obtained from each method and the true underlying signal $\tilde{\mathbf{X}}$. The fitted matrices from PCA and rPCA are obtained considering the true number of underlying dimensions as known. The SURE method is performed with the true noise variance as in Candès et al (2012). Results of the simulation study are gathered in Table 1.

First, rPCA outperforms both PCA and the SURE method in almost all situations. As expected, the MSE obtained by PCA and rPCA are roughly of the same order of magnitude when the SNR is high (SNR = 4), as illustrated in rows number 1 or 13, whereas rPCA outperforms PCA when data are noisy (SNR = 0.8) as in rows number 11 or 23. The differences between rPCA and PCA are also more critical when the ratio (d_1/d_2) is high than when the eigenvalues are equal. When (d_1/d_2) is large, the signal is concentrated on the first dimension whereas it is scattered in more dimensions when the ratio is smaller. Consequently, the same amount of noise has a greater impact on the second dimension in the first case. This may increase the advantage of rPCA which tends to reduce the impact of noise.

The main characteristic of the SURE method observed in all simulations is that it gives particularly good results when the data are very noisy. Consequently, the results are satisfactory when SNR = 0.8, particularly when the number of underlying dimensions is high (rows number 11, 23 and 35 for instance). This behaviour can be explained by the fact that the same amount of signal is more impacted by the noise if the signal is scattered on many

Table 1: Mean Squared Error (and its standard deviation) between $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$ for PCA, rPCA and SURE method over 500 simulations. Results are given for different numbers of individuals (n), numbers of variables (p), numbers of underlying dimensions (S), signal-to-noise ratios (SNR) and ratios of the first eigenvalue on the second eigenvalue (d_1/d_2).

	n	p	S	SNR	(d_1/d_2)	$MSE(\hat{\mathbf{X}}^{\text{PCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\text{rPCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\text{SURE}}, \tilde{\mathbf{X}})$
1	100	20	2	4	4	4.22E-04 (1.69E-06)	**4.22E-04** (1.69E-06)	8.17E-04 (2.67E-06)
2	100	20	2	4	1	4.21E-04 (1.75E-06)	**4.21E-04** (1.75E-06)	8.26E-04 (2.89E-06)
3	100	20	2	1	4	1.26E-01 (5.29E-04)	**1.08E-01** (4.56E-04)	1.60E-01 (6.15E-04)
4	100	20	2	1	1	1.23E-01 (5.05E-04)	**1.11E-01** (4.61E-04)	1.69E-01 (6.28E-04)
5	100	20	2	0.8	4	3.34E-01 (1.38E-03)	**2.40E-01** (9.90E-04)	3.10E-01 (1.05E-03)
6	100	20	2	0.8	1	3.12E-01 (1.38E-03)	**2.45E-01** (1.10E-03)	3.32E-01 (1.22E-03)
7	100	20	4	4	4	8.25E-04 (2.39E-06)	**8.24E-04** (2.38E-06)	1.42E-03 (3.54E-06)
8	100	20	4	4	1	8.26E-04 (2.38E-06)	**8.25E-04** (2.38E-06)	1.43E-03 (3.48E-06)
9	100	20	4	1	4	2.60E-01 (8.44E-04)	**1.96E-01** (6.51E-04)	2.43E-01 (6.84E-04)
10	100	20	4	1	1	2.47E-01 (7.16E-04)	**2.04E-01** (5.99E-04)	2.62E-01 (6.94E-04)
11	100	20	4	0.8	4	7.41E-01 (2.69E-03)	**4.27E-01** (1.53E-03)	4.36E-01 (1.11E-03)
12	100	20	4	0.8	1	6.68E-01 (2.02E-03)	**4.40E-01** (1.40E-03)	4.83E-01 (1.33E-03)
13	50	50	2	4	4	2.81E-04 (1.32E-06)	**2.81E-04** (1.32E-06)	5.95E-04 (2.24E-06)
14	50	50	2	4	1	2.79E-04 (1.24E-06)	**2.79E-04** (1.24E-06)	5.93E-04 (2.21E-06)
15	50	50	2	1	4	8.48E-02 (4.09E-04)	**7.82E-02** (3.85E-04)	1.26E-01 (4.97E-04)
16	50	50	2	1	1	8.21E-02 (3.87E-04)	**7.77E-02** (3.70E-04)	1.31E-01 (5.08E-04)
17	50	50	2	0.8	4	2.30E-01 (1.12E-03)	**1.93E-01** (9.64E-04)	2.55E-01 (1.01E-03)
18	50	50	2	0.8	1	2.14E-01 (9.58E-04)	**1.89E-01** (8.57E-04)	2.73E-01 (1.07E-03)
19	50	50	4	4	4	5.48E-04 (1.84E-06)	**5.48E-04** (1.84E-06)	1.04E-03 (2.82E-06)
20	50	50	4	4	1	5.46E-04 (1.76E-06)	**5.46E-04** (1.76E-06)	1.04E-03 (2.79E-06)
21	50	50	4	1	4	1.75E-01 (6.21E-04)	**1.53E-01** (5.54E-04)	2.00E-01 (5.79E-04)
22	50	50	4	1	1	1.68E-01 (5.49E-04)	**1.52E-01** (5.08E-04)	2.09E-01 (6.04E-04)
23	50	50	4	0.8	4	5.07E-01 (1.90E-03)	3.87E-01 (1.53E-03)	**3.85E-01** (1.12E-03)
24	50	50	4	0.8	1	4.67E-01 (1.62E-03)	**3.76E-01** (1.38E-03)	4.13E-01 (1.23E-03)
25	20	100	2	4	4	4.22E-04 (1.72E-06)	**4.22E-04** (1.72E-06)	8.15E-04 (2.80E-06)
26	20	100	2	4	1	4.21E-04 (1.69E-06)	**4.20E-04** (1.70E-06)	8.20E-04 (2.89E-06)
27	20	100	2	1	4	1.25E-01 (5.35E-04)	**1.06E-01** (4.53E-04)	1.57E-01 (5.83E-04)
28	20	100	2	1	1	1.22E-01 (5.28E-04)	**1.10E-01** (4.76E-04)	1.67E-01 (6.20E-04)
29	20	100	2	0.8	4	3.30E-01 (1.43E-03)	**2.35E-01** (1.03E-03)	3.06E-01 (1.13E-03)
30	20	100	2	0.8	1	3.18E-01 (1.30E-03)	**2.50E-01** (1.03E-03)	3.34E-01 (1.25E-03)
31	20	100	4	4	4	8.28E-04 (2.38E-06)	**8.27E-04** (2.39E-06)	1.41E-03 (3.64E-06)
32	20	100	4	4	1	8.29E-04 (2.58E-06)	**8.28E-04** (2.58E-06)	1.42E-03 (3.68E-06)
33	20	100	4	1	4	2.55E-01 (7.59E-04)	**1.97E-01** (5.92E-04)	2.45E-01 (6.47E-04)
34	20	100	4	1	1	2.48E-01 (7.45E-04)	**2.04E-01** (6.20E-04)	2.60E-01 (6.91E-04)
35	20	100	4	0.8	4	7.13E-01 (2.55E-03)	**4.15E-01** (1.47E-03)	4.37E-01 (1.19E-03)
36	20	100	4	0.8	1	6.66E-01 (2.01E-03)	**4.34E-01** (1.31E-03)	4.78E-01 (1.24E-03)

dimensions than if it is concentrated on few dimensions. This remark highlights the fact that the SNR is not necessarily a good measure of the level of noise in a data set. In addition, the results of the SURE method are quite poor when the SNR is high. This can be explained by the fact that the SURE method takes into account too many dimensions (since all the singular values which are higher than the threshold λ are kept) in the estimation of $\hat{\mathbf{X}}^{\text{SURE}}$. For example, with $n = 100$, $p = 20$, $S = 2$, $SNR = 4$ and $(d_1/d_2) = 4$ (first row), the SURE method considers between 9 and 13 dimensions to estimate $\hat{\mathbf{X}}^{\text{SURE}}$.

Finally, the behaviour regarding the ratio (n/p) is worth noting of. The MSEs are in the same order of magnitude for $(n/p) = 0.2$ and $(n/p) = 5$ and are much smaller for $(n/p) = 1$ for all the methods. The issue of dimensionality does not occur only when the number of variables is much larger than the number individuals. Rather, difficulties arise when one mode (n or p) is larger than the other one, which can be explained by the bilinear form of the model.

3.2 Simulations from Candès et al (2012)

Regularised PCA is also assessed using the simulations from Candès et al (2012). Simulated matrices of size 200×500 were drawn with 4 SNR (0.5, 1, 2 and 4) and 2 numbers of underlying dimensions (10, 100).

Results for the SURE method (Table 2) are in agreement with the results obtained by Candès et al (2012). As in the first simulation study (section 3.1), rPCA outperforms both PCA and the SURE method in almost all cases. However, the SURE method provides better results than rPCA when the number of underlying dimensions S is high ($S = 100$) and the SNR is small (SNR = 1, 0.5). This is in agreement with the previous comments highlighting the ability of the SURE method to handle noisy situations. Nevertheless, we note that when the SNR is equal to 0.5, rPCA is performed with the "true" number of underlying dimensions (100). However, if we estimate the number of underlying dimensions on these data with one of the available methods in the literature (Jolliffe, 2002), all the methods select 0 dimensions. Indeed, the data are so noisy that the signal is nearly lost. Results obtained with rPCA, using 0 dimensions results in estimating all the values of $\hat{\mathbf{X}}^{\text{rPCA}}$ by 0 which corresponds to an MSE equal to 1. In this case, considering 0 dimensions in rPCA leads to a lower MSE than taking into account 100 dimensions (MSE = 1.48), but it is still higher than the MSE of the SURE method (0.85).

The R (R Core Team, 2012) code to perform all the simulations is available on request.

3.3 Recovery of the graphical outputs

Because rPCA better recovers the signal, it produces graphical outputs (individual and variable representations) closer to the outputs obtained from $\tilde{\mathbf{X}}$. We illustrate this point on a simple data set with 100 individuals, 20 variables, 2 underlying dimensions, $(d_1/d_2) = 4$ and a SNR equal to 0.8 (row 5 of Table 1). Figure 2 provides the true individuals representation obtained from $\tilde{\mathbf{X}}$ (top left) as well as the representations obtained by PCA (top right), rPCA (bottom left) and the SURE method (bottom right). The cloud associated with PCA has a higher variability than the cloud associated with rPCA which is tightened around the origin. The effect of regularisation is stronger on the second axis than on the first one, which is expected because of the regularisation term. For instance, the individuals 82 and 59, which have small coordinates on the second axis in PCA are brought closer to the origin in the representation obtained by rPCA which is more in agreement with the true configuration. The cloud associated with the SURE method is tightened around the origin on the first axis and even more so on the second one, which is also expected because of the regularisation term. However the global variance of the SURE representation, which is reflected by the variability, is clearly lower than the variance of the true signal. Therefore, the global shape of the cloud of rPCA is the closest to the true one and thus rPCA successfully recovers the distances between individuals.

Figure 3 provides the corresponding representations for the variables. The link between the variables which have high coordinates on the first and the second axis of the PCA of $\mathbf{X}$ is reinforced in rPCA. This is consistent with the representation of $\tilde{\mathbf{X}}$. For instance, variables 9 and 7 which are correlated to 1 in $\tilde{\mathbf{X}}$ are not very linked in the PCA representation (correlation equal to 0.68) whereas their correlation equals 0.81 in the rPCA representation and 0.82 in the SURE representation. On the contrary, variables 20 and 7, orthogonal in $\tilde{\mathbf{X}}$, have rather high coordinates, in absolute value, on the second axis in the PCA representation (correlation equal to -0.60). Their link is slightly weakened in the rPCA representation (correlation equal to -0.53) and in the SURE representation (correlation equal to -0.51). In addition, all the variables are generated with a variance equal to 1. The variances are over-estimated in the PCA representation and under-estimated in the SURE represen-

Table 2: Mean Squared Error (and its standard deviation) between $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$ for PCA, regularised PCA (rPCA) and SURE method over 100 simulations. Results are given for $n = 200$ individuals, $p = 500$ variables, different numbers of underlying dimensions (S) and signal-to-noise ratios (SNR).

S	SNR	$MSE(\hat{\mathbf{X}}^{\text{PCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\text{rPCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\text{SURE}}, \tilde{\mathbf{X}})$
10	4	4.31E-03 (7.96E-07)	**4.29E-03** (7.91E-07)	8.74E-03 (1.15E-06)
10	2	1.74E-02 (2.84E-06)	**1.71E-02** (2.81E-06)	3.29E-02 (4.68E-06)
10	1	7.16E-02 (1.25E-05)	**6.75E-02** (1.15E-05)	1.16E-01 (1.59E-05)
10	0.5	3.19E-01 (5.44E-05)	**2.57E-01** (4.85E-05)	3.53E-01 (5.42E-05)
100	4	3.79E-02 (2.02E-06)	**3.69E-02** (1.93E-06)	4.50E-02 (2.12E-06)
100	2	1.58E-01 (8.99E-06)	**1.41E-01** (7.98E-06)	1.56E-01 (8.15E-06)
100	1	7.29E-01 (4.84E-05)	4.91E-01 (2.96E-05)	**4.48E-01** (2.26E-05)
100	0.5	3.16E+00 (1.65E-04)	1.48E+00 (1.12E-04)	**8.52E-01** (3.07E-05)

tation, particularly for the variables which are highly linked to the second axis. The best compromise for the variances is provided by rPCA. Therefore, rPCA successfully recovers the variances and the covariances of the variables.

This example shows that rPCA is a good method to recover the distances between individuals as well as the links between variables. This property of preserving distances is crucial in clustering for instance, as we will show in the applications (section 4).

4 Applications

4.1 Transcriptome profiling

Regularised PCA is applied to a real data set (Désert et al, 2008) which consists of a collection of 12664 gene expressions in 27 chickens submitted to 4 nutritional statuses: continuously fed (N), fasting for 16 hours (F16), fasting for 16 hours then refed for 5 hours (F16R5), fasting for 16 hours then refed for 16 hours (F16R16).

Since there are 4 nutritional statuses, 3 dimensions are considered. We expect the first three principal components to represent the between-class variability, whereas the following components represent the within-class variability which is less of interest. Figure 4 shows the individual representations obtained by PCA (top left), rPCA (top right) and the SURE method (bottom left). To better highlight the effect of regularisation, dimensions 1 and 3 are presented. The first dimension of PCA, rPCA and the SURE method order the nutritional statuses from the continuously fed chickens (on the left) to the fasting chickens (on the right). Chickens N.4 and F16R5.1, which have high coordinates in absolute value on the third axis of PCA, are brought closer to the other chickens submitted to the same status in the rPCA representation and in the SURE representation. In addition, chickens N.1 and F16.4, which have high coordinates on the first axis are brought closer to the origin in the SURE representation. Despite these differences, the impact of the regularisation on the graphical outputs appears to be small.

The representation obtained after a sparse PCA (sPCA) method (Witten et al, 2009) implemented in the R package `PMA` (Witten et al, 2011) is also provided (bottom right). Indeed, it is very common to use sparse methods on this kind of data (Zou et al, 2006). The basic assumptions for the development of sPCA is that PCA provides principal components that are linear combinations of the original variables which may lead to difficulties during the interpretation especially when the number of variables is very large. Loadings obtained via sPCA are indeed sparse, meaning they contain many 0 elements and therefore select only a few variables. The representation stemming

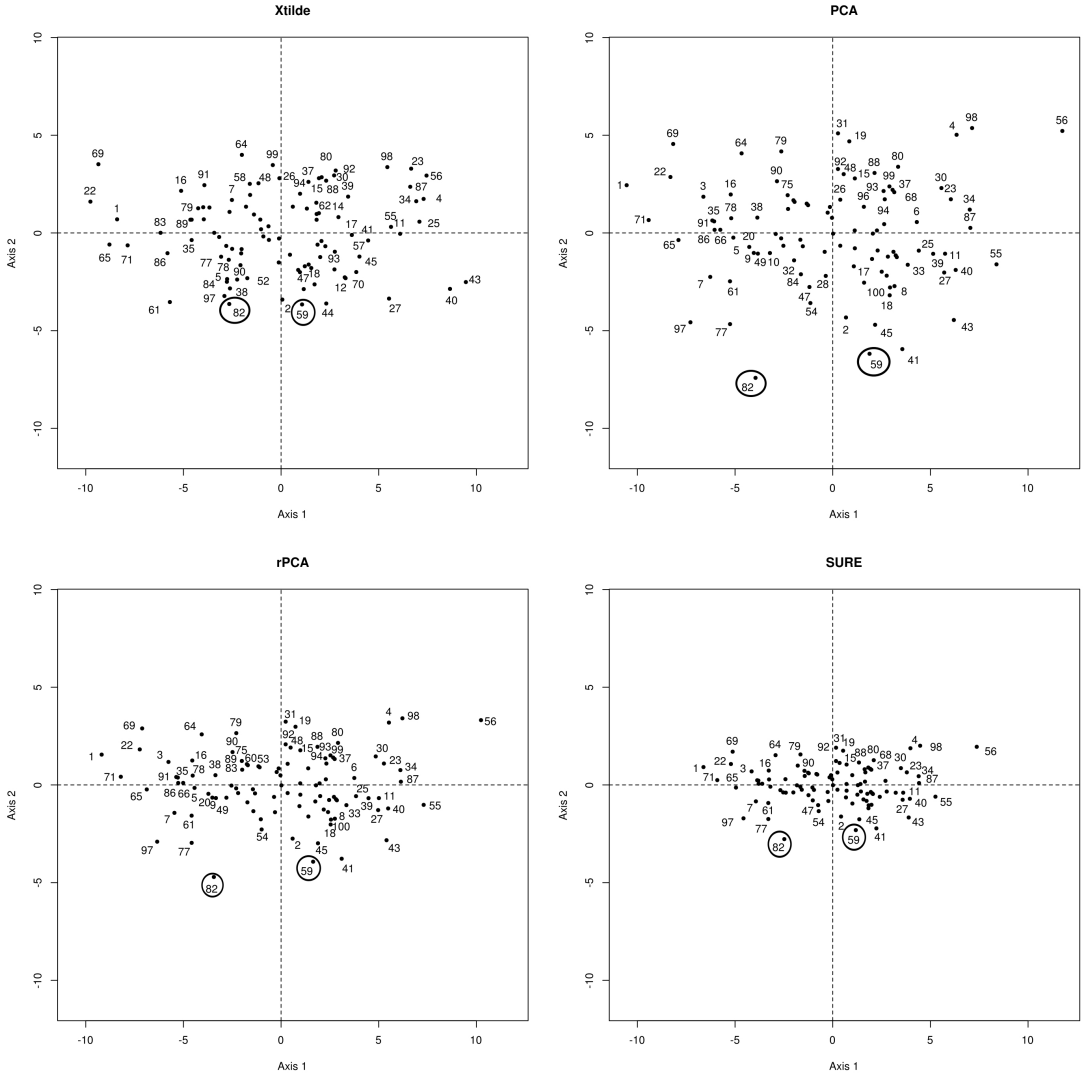


Fig. 2: Individual representations of $\tilde{\mathbf{X}}$ (top left), of the PCA of $\mathbf{X}$ (top right), of the rPCA of $\mathbf{X}$ (bottom left) and of the SURE method applied to $\mathbf{X}$ (bottom right) for a data set with $n = 100$, $p = 20$, $S = 2$, $(d_1/d_2) = 4$ and SNR= 0.8.

from sPCA is quite different from the other representations; in particular the clusters of F16R5 and of F16 chickens are less clearly differentiated.

It is customary to complement principal components methods with double clustering in order to simultaneously cluster the chickens and the genes and to represent the results using heatmaps (Eisen et al, 1998). The heatmap clustering is applied to the matrices $\hat{\mathbf{X}}$ obtained by the different methods (Figure 5). Because rPCA modifies the distances between chickens as well as the covariances between genes, the rPCA heatmap will differ from the PCA heatmap. The rPCA heatmap (Figure 5b) is much more appropriate than the PCA heatmap (Figure 5a). Indeed, the chickens undergoing 16 hours

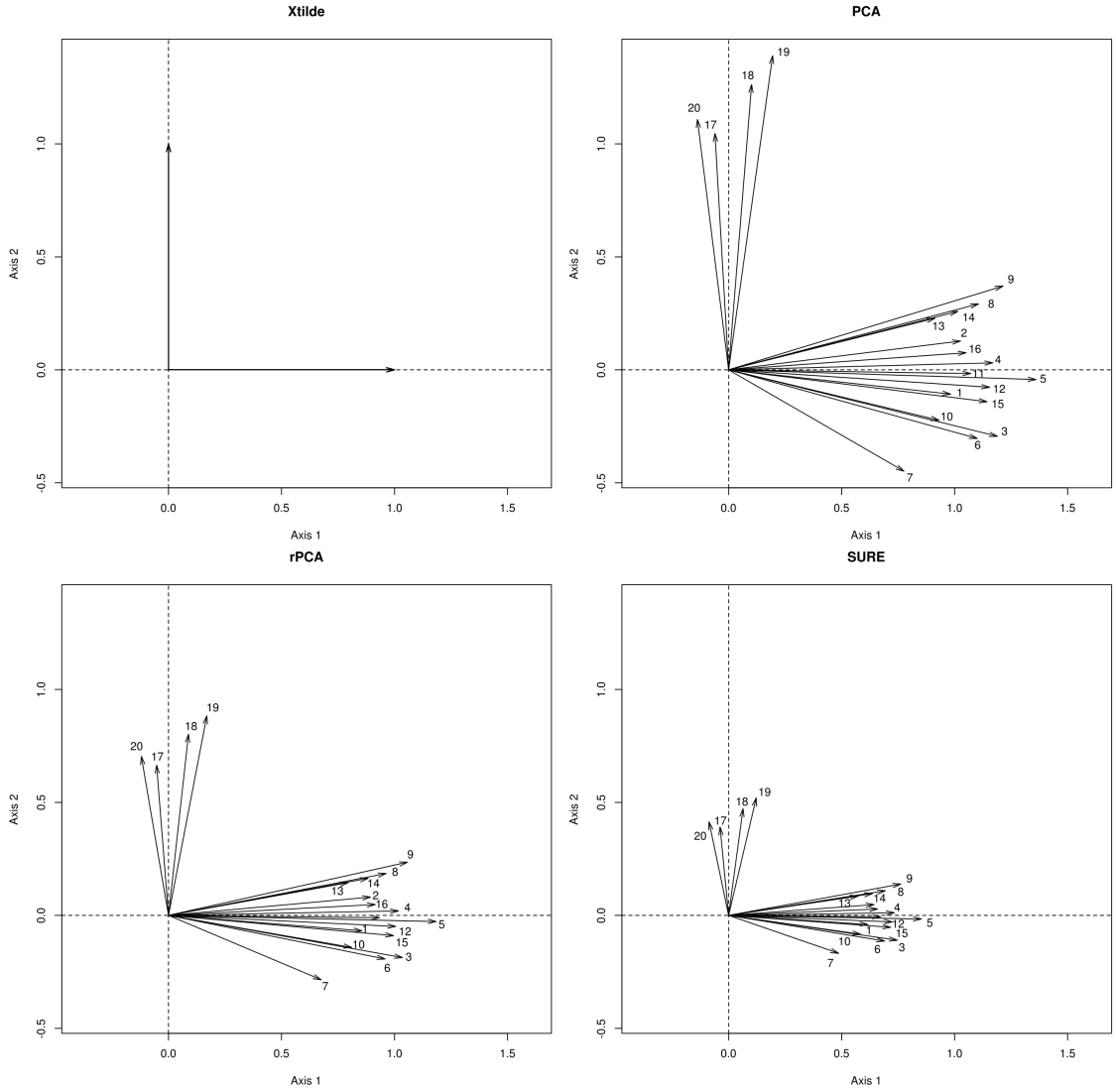


Fig. 3: Variable representations of the PCA of $\tilde{\mathbf{X}}$ (top left), the PCA of $\mathbf{X}$ (top right), the rPCA of $\mathbf{X}$ (bottom left) and the SURE method applied to $\mathbf{X}$ (bottom right) for an example of data set with $n = 100$, $p = 20$, $S = 2$, $(d_1/d_2) = 4$ and SNR= 0.8.

of fasting are separated into two sub-clusters in the PCA heatmap separated by the chickens F16R5.1, F16R16.3 and F16R16.4, whereas they are well-clustered in the rPCA heatmap. Similarly chickens F16R5 are agglomerated in the PCA heatmap except for chickens F16R5.1 and F16R5.3, whereas they are well-clustered in the rPCA heatmap. Finally, the F16R16 chickens are more scattered in both representations. However in rPCA, this can be interpreted as some of the chickens, having fully recovered from the fasting period, are mixed with continuously fed chickens, and some having not fully recovered are mixed with F16R5 chickens: the large majority of F16R16 chickens are agglomerated and mixed with N.6 and N.7, and chicken F16R16.1 is mixed with F16R5 chickens. It is not the case for PCA,

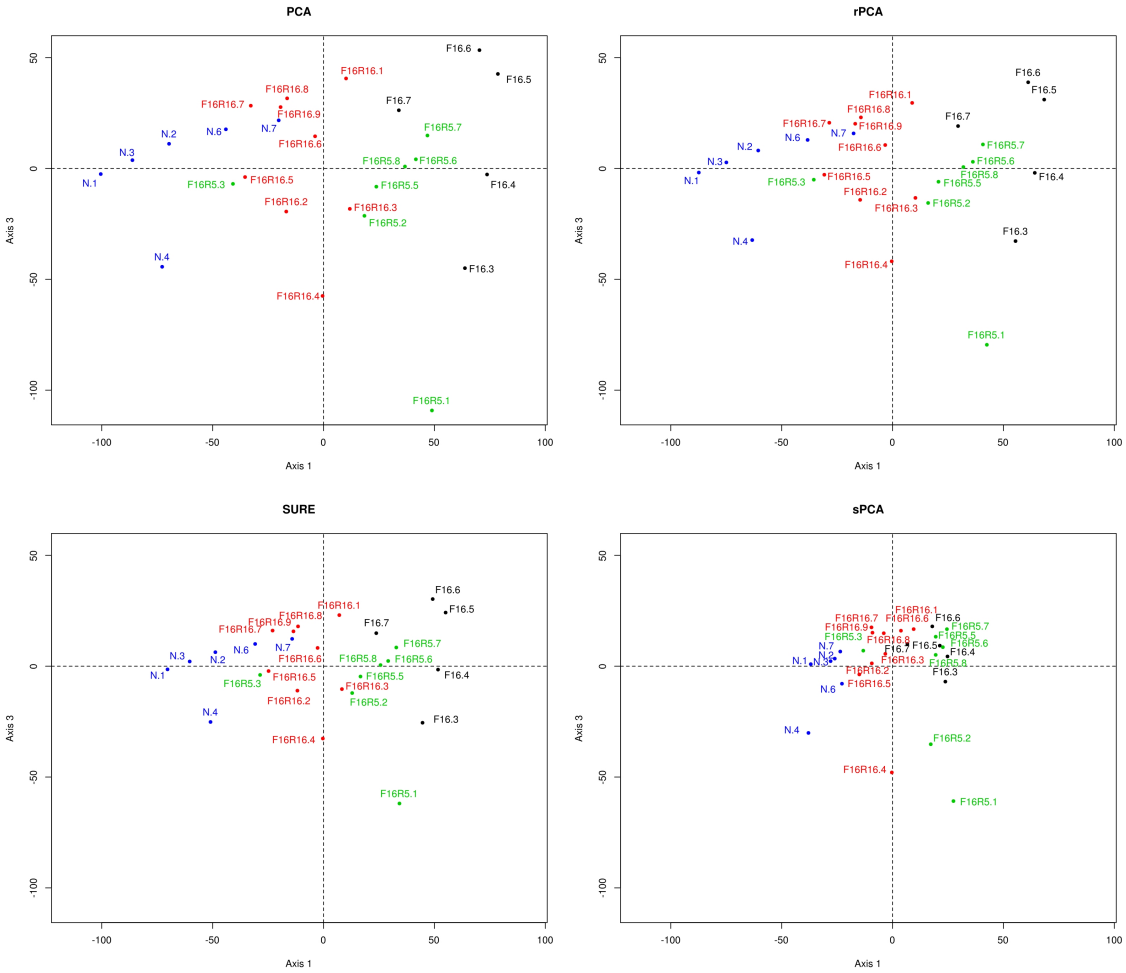
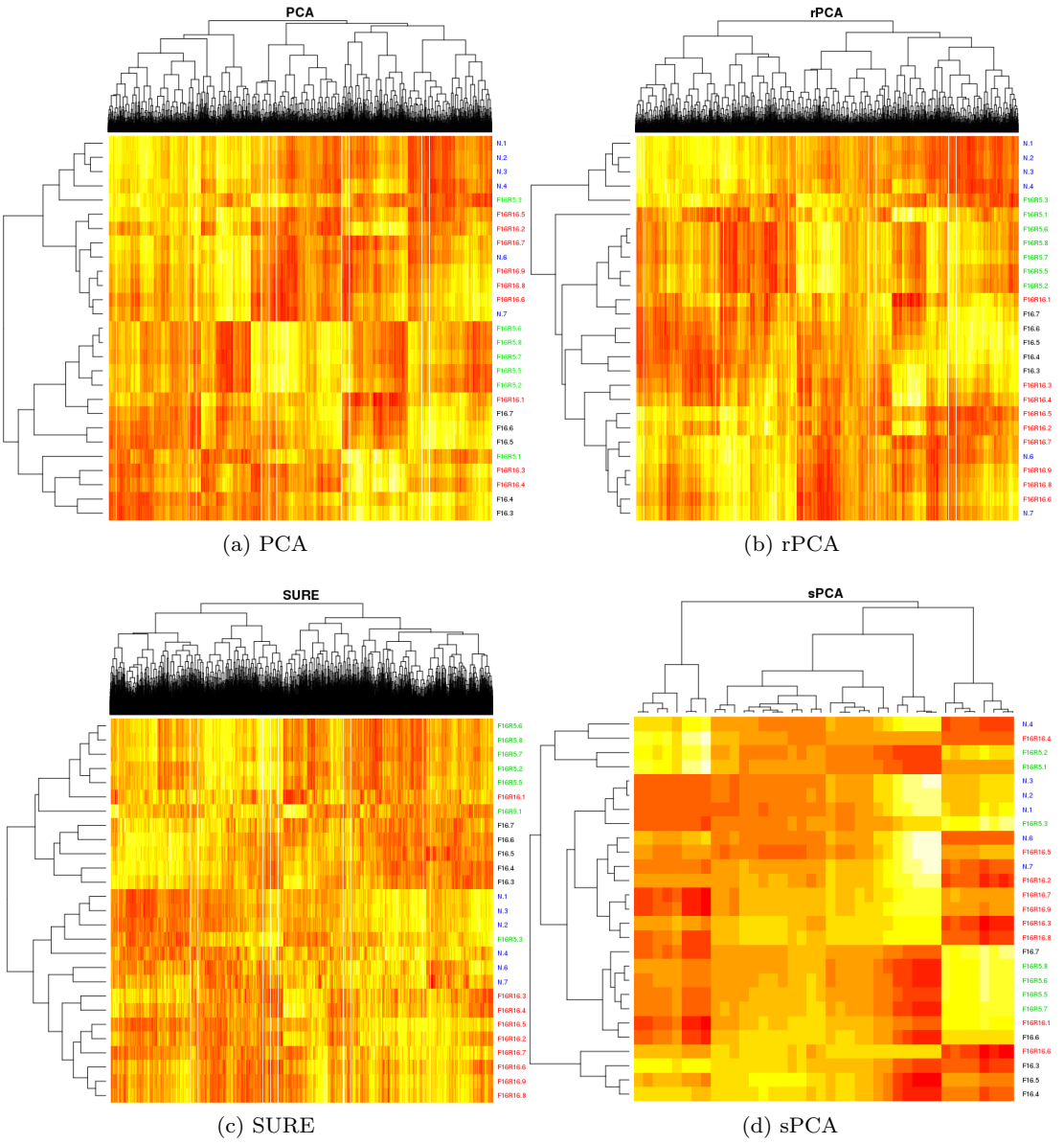


Fig. 4: Representation of the individuals on dimensions 1 and 3 of the PCA (top left), the rPCA (top right), the SURE method (bottom left) and sPCA (bottom right) of the transcriptome profiling data. Individuals are coloured according to the nutritional statuses.

where the F16R16 chickens are mixed with chickens submitted to all the other nutritional statuses. The conclusions concerning the SURE heatmap (Figure 5c) are similar to the conclusions drawn from rPCA. The 4 clusters corresponding to the 4 nutritional statuses are well-defined. However, chicken F16R5.3 is clustered with the N chickens. In addition, the global contrasts are weaker in the SURE heatmap than in the rPCA heatmap. The heatmap stemming from sPCA (Figure 5d) seems to be easier to interpret since there are more contrasts. This is due to the drastic selection of the genes (43 genes were selected among the 12664 genes of the data set). However none of the chicken clusters is clearly defined.

We will not dwell on the interpretation of the gene expressions in the heatmap; however, if the chicken clustering is coherent, the gene clustering is expected to be more coherent as well.

In this example, the impact of regularisation on the graphical representations is not obvious, but the effect of regularisation is crucial



(a) PCA (b) rPCA

(c) SURE (d) sPCA

Fig. 5: Heatmaps associated with the analysis of the transcriptomic data set. The data sets used to perform the heatmaps are the fitted matrices stemming from PCA (a), rPCA (b), the SURE method (c) and sPCA (d).

to the results of the clustering. This can be explained by the ability of rPCA to denoise data. Such a denoising property can also be useful when dealing with images as illustrated in the next section.

4.2 Image denoising

We consider the PINCAT numerical Phantom data from Sharif and Bresler (2007) analysed in Candès et al (2012) providing a signal with

complex values. The PINCAT data simulate a first-pass myocardial perfusion real-time magnetic resonance imaging series, comprising 50 images, one for each time. To compare the performances of PCA, rPCA and the SURE method, 100 data sets are generated by adding a complex iid Gaussian noise, with a standard deviation equal to 30, to the PINCAT image data. The original image data are then considered as the true (noise-free) images. PCA and rPCA are performed assuming 20 underlying dimensions. This number was chosen empirically and we verified that using slightly more or fewer dimensions does not greatly impact the results. The SURE method is performed by taking into account the true noise standard deviation which is equal to 30. The methods are then evaluated by computing the MSE over the 100 simulations. The MSE are equal to 814.26, 598.17 and 727.33 respectively for PCA, rPCA and the SURE method. Consequently, rPCA outperforms both PCA and the SURE method in terms of MSE.

In addition, Figure 6 presents a comparison on one simulation of PCA, rPCA and the SURE method. Similarly to Candès et al (2012), we present 3 frames from the PINCAT data (early, middle and late times) for the true image data, the noisy image data, and the image data resulting from denoising by PCA, rPCA and SURE. All three methods are clearly efficient to reduce the noise; however, the SURE method and rPCA provide images with more contrast than the images provided by PCA. Since rPCA has lower MSE it provides images with a higher degree of noise reduction.

In addition, we can consider the worst-case absolute error through time (Figure 7), which is the highest residual error for each pixel at any time. The SURE method has a particularly high residual error in the area near the myocardium which is an area of high motion. The residual error is globally lower for rPCA than for SURE, and it is overall lower in the myocardium area.

Therefore, rPCA is a very promising method to denoise image data.

Conclusion

When data can be seen as a true signal corrupted by error, PCA does not provide the best recovery of the underlying signal. Shrinking the singular values improves the estimation of the underlying structure especially when data are noisy. Soft thresholding is one of the most popular strategies and consists in linearly shrinking the singular values. The regularised PCA suggested in this paper applies a nonlinear transformation of the singular values associated with a hard thresholding rule. The regularised term is analytically derived from the MSE using asymptotic results from nonlinear regression models or using Bayesian considerations. In the simulations, rPCA outperforms the SURE method in most situations. We showed in particular that rPCA can be used beneficially prior to clustering (of individuals and/or variables) or in image denoising. In addition, rPCA allows improvement on the graphical representations in an exploratory framework. In this framework, it is worth quoting the work of Takane and Hwang (2006) and Hwang et al (2009) who suggested a regularised version of multiple correspondence analysis, which also improves the graphical representations.

Regularised PCA requires a tuning parameter which is the number of underlying dimensions. Many methods (Jolliffe, 2002) are available in the literature to select this parameter. However, it is still a difficult problem and an active research area. A classical statement is the following: if the selected number of dimensions is smaller than the rank S of the signal, some of the relevant information is lost and, in our situation, this results in overestimating the noise variance. On the contrary, selecting more than S dimensions appears preferable because all the signal is taken into account even if the noise variance is underestimated. However, in case of very noisy data, the signal is overwhelmed by the noise and is nearly lost. In such a case, it is better to select a number of dimensions smaller than S. This strat-

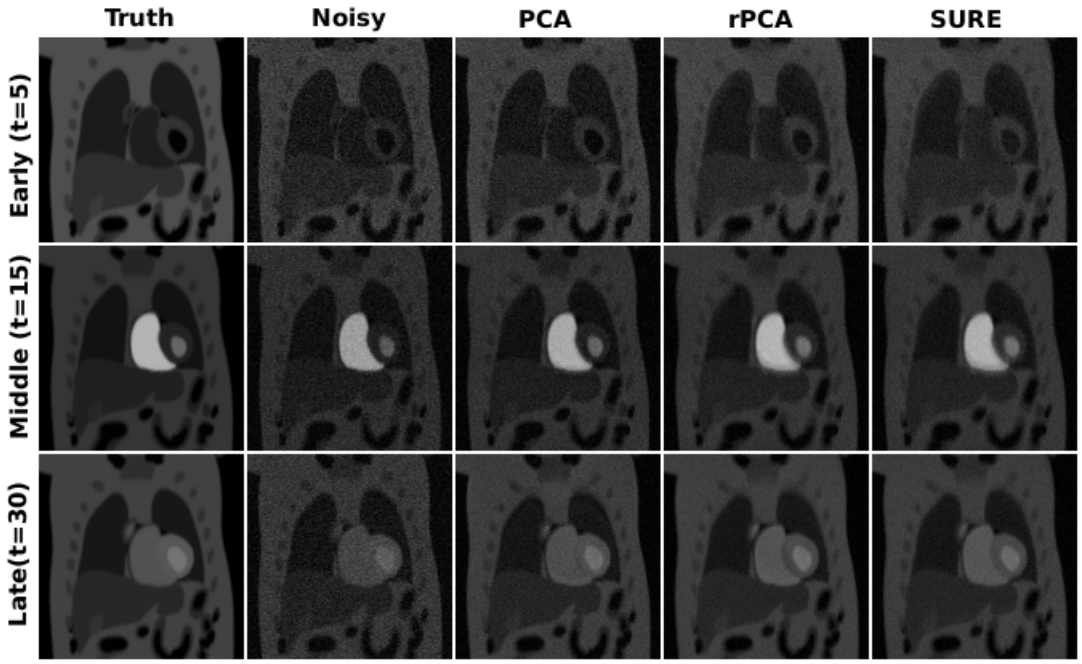


Fig. 6: Representation of frames from the PINCAT data at 3 times (early, middle and late) of the true images, the noisy images and the image estimations resulting from PCA, rPCA and SURE.

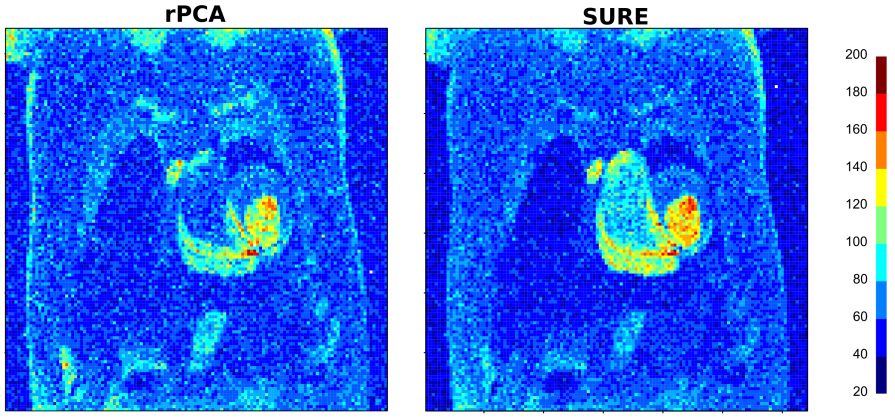


Fig. 7: Worst-case absolute error through time of the image estimations by rPCA and SURE.

egy is a way to regularise more which is acceptable when data are very noisy. In practice, we use a cross-validation strategy (Josse and Husson, 2011) which behaves desirably in our simulations (that is, to find the true number of dimensions when the signal-to-noise ratio is large, and to find a smaller number when the signal-to-noise ratio is small).

References

Bartholomew D (1987) Latent Variable Models and Factor Analysis. Charles Griffin and

Co Ltd

Candès EJ, Tao T (2009) The power of convex relaxation: Near-optimal matrix completion. IEEE Trans Inform Theory 56(5):2053–2080

Candès EJ, Sing-Long CA, Trzasko JD (2012) Unbiased risk estimates for singular value thresholding and spectral estimators, (Submitted)

Caussinus H (1986) Models and uses of principal component analysis (with discussion), DSWO Press, p 149–178

Chikuse Y (2003) Statistics on Special Manifolds. Springer

Cornelius P, Crossa J (1999) Prediction assessment of shrinkage estimators of multiplicative models for multi-environment cultivar trials. Crop Science 39:998–1009

Denis JB, Gower JC (1994) Asymptotic covariances for the parameters of biadditive models. Utilitas Mathematica pp 193–205

Denis JB, Gower JC (1996) Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. Journal of the Royal Statistical Society Series C (Applied Statistics) 45:479–493

Denis JB, Pázman A (1999) Bias of least squares estimators in nonlinear regression models with constraints. part ii: Biadditive models. Applications of Mathematics 44:359–374

Désert C, Duclos M, Blavy P, Lecerf F, Moreews F, Klopp C, Aubry M, Herault F, Le Roy P, Berri C, Douaire M, Diot C, S L (2008) Transcriptome profiling of the feeding-to-fasting transition in chicken liver. BMC Genomics

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95(25):14,863–14,868

Gower JC, Dijksterhuis GB (2004) Procrustes Problems. Oxford University Press, USA

Greenacre MJ (2010) Biplots in Practice. BBVA Fundation

Hastie TJ, Tibshirani RJ, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second edition. Springer

Hoff PD (2007) Model averaging and dimension selection for the singular value decomposition. J Amer Statist Assoc 102(478):674–685

Hoff PD (2009) Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. Journal of Computational and Graphical Statistics 18(2):438–456

Husson F, Le S, Pages J (2010) Exploratory Multivariate Analysis by Example Using R, 1st edn. CRC Press

Hwang H, Tomiuk M, Takane Y (2009) Correspondence Analysis, Multiple Correspondence Analysis and Recent Developments, Sage Publications, pp 243–263

Jolliffe I (2002) Principal Component Analysis. Springer Series in Statistics

Josse J, Husson F (2011) Selecting the number of components in pca using cross-validation approximations. Computational Statististics and Data Analysis 56:1869–1879

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research 99:2287–2322

Papadopoulo T, Lourakis MIA (2000) Estimating the jacobian of the singular value decomposition: Theory and applications. In: In Proceedings of the European Conference on Computer Vision, ECCV00, Springer, pp 554–570

R Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, ISBN 3-900051-07-0

Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. Statistical Science 6(1):15–32

Roweis S (1998) Em algorithms for pca and spca. In: Advances in Neural Information Processing Systems, MIT Press, pp 626–632

Rubin DB, Thayer DT (1982) EM algorithms for ML factor analysis. Psychometrika 47(1):69–76

Sharif B, Bresler Y (2007) Physiologically improved NCAT phantom (PINCAT) enables in-silico study of the effects of beat-to-beat variability on cardiac MR. In: Proceedings of the Annual Meeting of ISMRM, Berlin, p 3418

Takane Y, Hwang H (2006) Regularized multiple correspondence analysis, Boca Raton, FL: Chapman & Hall/CRC, pp 259–279

Tipping M, Bishop C (1999) Probabilistic principal component analysis. Journal of the Royal Statistical Society (Series B) 61:611–622

Witten D, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10:515–534

Witten D, Tibshirani R, Gross S, Narasimhan B (2011) PMA: Penalized Multivariate Analysis. URL `http://CRAN.R-project.org/package=PMA`, r package version 1.0.8

Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. Journal of Computational and Graphical Statistics 15(2):265–286

CHAPITRE 3

INTÉGRATION D'INFORMATION BIOLOGIQUE

Dans ce chapitre, nous proposons un nouvel algorithme de clustering de gènes basé sur l'intégration d'information biologique dans l'analyse exploratoire de données d'expression. Le principe de cette méthodologie repose sur le constat que la coexpression entre deux gènes et par conséquent les clusters de gènes coexprimés sont d'interprétation délicate. C'est pourquoi nous proposons une nouvelle méthodologie de clustering de gènes basée sur la coexpression mais également sur l'association des gènes à des annotations fonctionnelles de type Gene Ontology. Ainsi, nous constituons des clusters de gènes qui sont à la fois coexprimés et associés à des annotations Gene Ontology similaires. Nous proposons également une procédure d'évaluation des clusters basée sur deux indicateurs associés à deux probabilités critiques qui mesurent respectivement la coexpression et l'homogénéité biologique des clusters.

Ce chapitre inclut l'article :

Verbanck, M., Lê, S., & Pagès, J. (2013). A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. *BMC Bioinformatics*, **14**, 42. (Highly Accessed)

Sommaire

1 Vers l'intégration d'information biologique

1.1 Démarche classique

Ainsi que nous l'avons établi section 1.4.3 chapitre 1, le cadre classique d'analyse des données transcriptomiques comporte une étape de visualisation des données très souvent couplée à une étape de clustering. La visualisation des données transcriptomiques repose classiquement sur des méthodes d'analyse multidimensionnelle exploratoire telles que l'ACP (chapitre 2). Cette étape est usuellement suivie (en ne prenant en compte qu'un certain nombre de composantes par exemple) d'une étape de clustering. Cette étape de clustering est donc réalisée sur une distance entre gènes dont le calcul n'est basé que sur les données d'expression. Nous pouvons également citer des exemples de traitements des données transcriptomiques qui consistent en une étape d'inférence de réseaux de régulation suivie d'une étape de clustering à partir du réseau. Dans ces deux cas, l'étape de clustering de gènes repose uniquement sur les données transcriptomiques, que ce soit directement comme dans les Heatmaps (Eisen *et al.*, 1998) ou indirectement suite à une étape d'inférence de réseaux de régulation (Zhang et Horvath, 2005). Les clusters ainsi obtenus sur la base des seules données transcriptomiques sont ensuite interprétés au moyen d'une information extérieure sur les gènes par l'intermédiaire de tests d'enrichissement, l'information extérieure étant tirée des bases de données de type Gene Ontology (section 1.4.4 chapitre 1).

La stratégie classique repose inévitablement sur des hypothèses qui ne sont pas nécessairement formulées mais que nous proposons d'expliciter. Premièrement, la caractérisation biologique de clusters de gènes coexprimés implique que des connexions biologiques existent de façon systématique entre gènes coexprimés. Deuxièmement, la caractérisation biologique s'appuie uniquement sur l'information biologique extérieure, ainsi il est attendu qu'une partie de cette information soit apparentée à l'expérience de l'étude.

1.2 Nécessité d'intégration d'information biologique

Comme nous l'avons déjà établi dans le chapitre 1, cette première hypothèse implicite qui consiste à affirmer que des connexions biologiques existent systématiquement entre gènes coexprimés est discutable (Raychaudhuri *et al.*, 2000). En effet, la coexpression des gènes est très délicate à interpréter car elle peut refléter des réalités biologiques très différentes. Ainsi nous sommes arrivés à la conclusion que la connaissance du seul transcriptome, en d'autres termes l'utilisation des seules données d'expression, n'est pas suffisante pour démêler les complexes relations qui existent entre gènes. Or, dans la démarche classique, le recours à de l'information extérieure sur les gènes dans l'interprétation finale des analyses peut être vu comme une façon de démêler toutes ces interprétations pos-

sibles. Cependant cette intégration, que nous pouvons qualifier de passive, d'information biologique dans l'analyse de données d'expression n'est sans doute pas suffisante. C'est pourquoi nous proposons de prendre en compte de façon active l'information biologique dans la constitution des clusters de gènes.

Ainsi, nous avons choisi de nous intéresser à de l'information extérieure sur les gènes sous forme de bases de données synthétiques telles que la base Gene Ontology dont nous rappelons le principe et les caractéristiques.

2 Gene Ontology

Depuis l'arrivée des technologies d'étude haut débit qui permettent de mesurer l'expression de milliers de gènes par exemple, l'interprétation des analyses nécessite d'avoir recours à de l'information synthétique. En effet, pour interpréter de tels ensembles de gènes, explorer toutes les connaissances et la bibliographie à propos de chaque gène semble inconcevable. Ainsi, sont apparues des bases de données synthétiques telles que la base de données Gene Ontology (GO). Le projet GO (Ashburner *et al.*, 2000) est un projet collaboratif et apporte une réponse à un problème de représentation des connaissances biologiques sur les gènes. La réponse proposée par le projet GO consiste en trois ontologies, une ontologie étant un ensemble structuré des termes et concepts décrivant les gènes : « Composant cellulaire », « Fonction moléculaire » et « Processus biologique ». Nous proposons de présenter dans un premier temps le principe détaillé et la structure de la base de données, puis nous en exposerons les limites.

2.1 Structure

Pour appréhender le concept de Gene Ontology, nous pouvons artificiellement distinguer deux composants : les trois ontologies avec leurs termes associés d'une part et les associations des gènes aux termes d'autre part.

2.1.1 Les ontologies

Intéressons-nous tout d'abord aux ontologies qui permettent une décomposition des connaissances sur les gènes. Nous pouvons considérer que les trois ontologies « Composant cellulaire », « Fonction moléculaire » et « Processus biologique », sont fixes et ne changent pas. Chaque ontologie repose sur un modèle de type graphe acyclique orienté (DAG). Les nœuds du DAG représentent des termes GO. Un terme GO est une connaissance sur les gènes qui peut être plus ou moins générale. Les branches du DAG représentent des inclusions entre les termes (figure 1). Ainsi un terme spécifique est inclus dans un terme plus général. Nous pouvons prendre l'exemple du terme « extracellular region part » qui représente un aspect plus spécifique du terme « extracellular region ». Le terme « extracellular region part » est donc inclus dans le terme « extracellular region », ce que matérialise la flèche entre ces deux termes. Précisons également que le terme GO

à la racine de chacune des trois ontologies est respectivement « Composant cellulaire », « Fonction moléculaire » et « Processus biologique ».

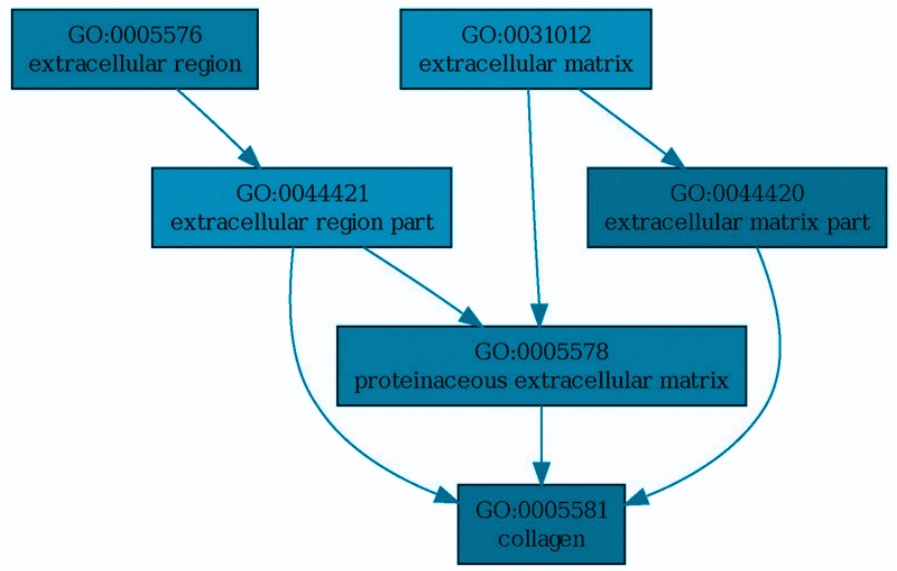


FIGURE 1 – Extrait de l'ontologie « Composant cellulaire » de Gene Ontology. (*source : http://omicslab.genetics.ac.cn/GOEAST/*)

Les trois ontologies et les termes associés représentent un modèle de connaissances représenté par la structure DAG qui est fixée et qui n'évolue pas. Ensuite les gènes sont décrétés *associés* aux termes GO.

2.1.2 Association des gènes aux termes

Les associations d'un gène aux termes GO sont décrétées en fonction des informations dont nous disposons sur le gène en question. Ces informations peuvent provenir de quatre sources principales :

- une vérification expérimentale reportée dans la littérature (par exemple un knock-out d'un gène qui consiste à éteindre le gène, à l'empêcher de s'exprimer afin de mettre en évidence ses fonctions)
- une déduction in silico (par exemple par alignement de séquences qui permet la détection d'un domaine spécifique dans la séquence d'ADN d'un gène et pourra conduire à associer au gène une certaine fonction)
- une déduction qui peut être indirectement dérivée des deux premiers types d'associations (par exemple des comparaisons inter-espèces)
- une origine inconnue (par exemple, si aucune information n'est disponible ou déductible à propos d'un gène, celui-ci peut être simplement associé aux termes racines)

Une fois les informations sur le gène identifiées, les associations sont décrétées selon quelques règles simples :

- le gène peut être associé à n'importe quel terme, quelle que soit la profondeur du terme, la profondeur étant le nombre de termes entre le terme en question et la racine dans l'architecture DAG
- si le gène est associé à un certain terme, ce gène est automatiquement associé à tous les termes parents du terme en question
- le gène peut être associé à deux termes « indépendants », c'est-à-dire qui ne présentent pas de relations directes

Nous insistons sur le fait que les associations d'un gène aux termes ne sont pas fixées et déterminées mais varient au fil des études et découvertes sur le gène.

2.2 Utilisation

Afin d'illustrer l'intérêt de cette information synthétique, nous proposons de considérer un exemple fictif et simplifié qui est schématisé figure 2. Prenons l'exemple fictif d'une réaction chimique de la cellule qui est la *synthèse du produit* à partir du produit A*. Cette réaction chimique fait intervenir trois enzymes (a, b, c) qui sont des protéines. Ces enzymes sont donc synthétisées suite à l'expression des trois gènes codant pour les trois enzymes (gène de l'enzyme a, gène de l'enzyme b et gène de l'enzyme c). Pour que ces trois gènes s'expriment, il est nécessaire d'avoir l'intervention d'un activateur de l'expression de ces trois gènes. Cet activateur est une protéine, produit de l'expression d'un gène activateur. Dans ce cas, il s'agit du gène A. Ainsi lorsque le gène A s'exprime, la protéine issue de l'expression du gène A active l'expression des gènes a, b et c. Il résulte de l'expression des gènes a, b et c, la production des trois enzymes correspondantes. La connaissance des phénomènes de régulation de l'expression des gènes impliqués dans la synthèse du synthèse du produit* à partir du produit A, peut être synthétisée dans l'ontologie « Processus Biologique » : les quatre gènes a, b, c et A sont ainsi associés au terme GO *synthèse du produit **.

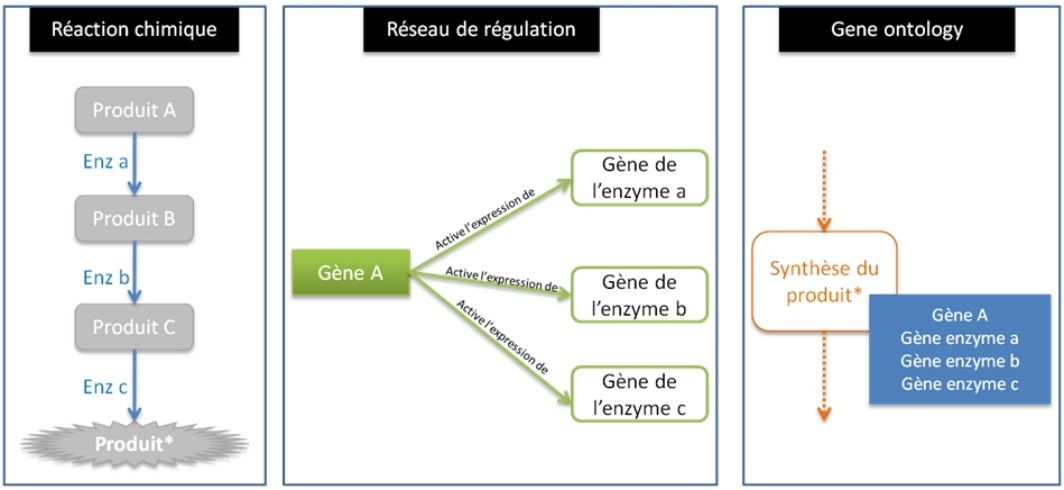


FIGURE 2 – Schéma de la structure des annotations Gene Ontology.

Par ailleurs, si nous considérons que dans nos données, les gènes A, a, b et c sont coexprimés, étant donné que ces quatre gènes sont associés au même terme GO, il est vraisemblable que ces gènes partagent un véritable lien biologique. Plus généralement, si deux gènes sont coexprimés et qu'ils sont associés à un grand nombre de termes GO communs, il est plus vraisemblable qu'il existe un véritable lien biologique entre ces deux gènes que s'ils ne partagent aucun terme GO.

A partir de cet exemple très simplifié, nous esquissons le principe et l'intérêt de l'intégration d'information biologique dans les données d'expression. Il existe cependant plusieurs limites aux annotations Gene Ontology que nous ne pourrons pas forcément prendre en compte mais dont il faut avoir connaissance.

2.3 Limites

Les bases de données synthétiques telles que la base Gene Ontology peuvent faire l'objet de critiques quant à la rigueur et la pertinence de l'information qu'elles fournissent. En effet, rassembler les connaissances et en trouver un système de représentation unique pour comparer les gènes est une tâche extrêmement ambitieuse. La base de données GO est donc nécessairement complétée via différentes sources d'information, ce qui peut induire des contradictions et des erreurs.

Pour un gène donné, les connaissances peuvent être partielles, et obtenues de façon indépendante (dans différentes expériences, conditions). Une des limites dont il faut être bien conscient pour le développement de méthodologies utilisant les annotations GO, réside dans le fait que la richesse et l'exhaustivité de la littérature sont très variables d'un gène à l'autre. En d'autres termes, le degré de connaissance et par conséquent le nombre d'associations aux termes GO est très variable d'un gène à l'autre.

En plus du fait que la littérature soit très variable d'un gène à l'autre, la littérature concernant un même gène est elle-même très hétérogène et peut concerner beaucoup d'aspects sans rapport du gène. Ces différents aspects sans rapport d'un gène sont expliqués par le fait qu'un même gène, qui agit à travers sa protéine, peut avoir plusieurs « fonctions ». D'une part, nous ne l'avons pas évoqué dans le chapitre 1 par soucis de simplification, mais il est possible qu'un gène soit traduit en plusieurs protéines. En effet, il peut exister une étape supplémentaire entre la transcription (copie de la séquence d'ADN en ARNm) et la traduction (synthèse de la protéine). L'ARNm peut subir des modifications qui conduiront à la synthèse de protéines légèrement différentes qui n'ont pas exactement les mêmes propriétés. Nous pouvons par exemple imaginer une protéine dont une partie de la séquence présente une charge négative, la fonction d'une telle protéine peut donc être de se fixer à des molécules portant une charge positive. Imaginons qu'une autre protéine soit traduite à partir d'un ARNm modifié et que cette protéine ne contienne plus la séquence portant une charge négative, celle-ci n'a donc pas la même fonction que la première protéine et ne peut pas se fixer à des molécules portant une charge positive. L'étape de modification de l'ARNm n'est pas systématique mais il est estimé que 70% des gènes peuvent subir cette étape, et l'on compte en moyenne 4 variants par gène.

Par ailleurs, les protéines peuvent s'associer entre elles. Par exemple dans la machinerie cellulaire de la transcription, les facteurs de transcription généraux se fixent à l'ARN polymérase (section 1.2 chapitre 1). Ainsi une même protéine peut avoir des fonctions différentes selon la protéine qui lui est associée.

Au vu de ces limites, il n'apparaît pas simple d'exploiter la base de données GO. Cependant, sachant que la quantité de connaissances est en perpétuelle augmentation, ce type de base de données synthétiques devient indispensable. De plus, nous pouvons penser que ces bases sont de plus en plus complètes et que leur qualité ne peut que s'améliorer avec les nouvelles connaissances disponibles. Ainsi, développer des outils qui utilisent les annotations GO peut être un choix judicieux : si la qualité de la base de données s'améliore, les outils ne peuvent que s'améliorer également. C'est une des raisons qui a motivé les choix méthodologiques que nous avons fait au cours de ce travail de recherche.

Nous pouvons donc penser, malgré les limites des annotations GO, que si deux gènes sont associés à un grand nombre de termes Gene Ontology communs, il existe plus vraisemblablement un véritable lien biologique (appartenance à un même réseau de régulation par exemple) entre ces deux gènes qu'entre deux gènes ne partageant aucun terme commun. Nous proposons donc de mettre au point une méthodologie de clustering de gènes basée sur l'intégration d'annotations GO dans les données d'expression.

3 Mise au point d'un algorithme d'intégration d'information biologique

Nous cherchons à intégrer de façon active de l'information sur les gènes, de type Gene Ontology, dans l'analyse de données d'expression.

3.1 Premiers essais d'intégration

Les premiers essais d'intégration d'information biologique dans l'analyse multidimensionnelle ont consisté à mettre en relation *a posteriori* l'information biologique avec les résultats des analyses (Busold *et al.*, 2005; Fagan *et al.*, 2007; Tayrac *et al.*, 2009). Dans ces premiers essais d'intégration, l'information biologique est utilisée comme une information de clusters sur les gènes. Cette information est projetée *a posteriori* en tant qu'éléments supplémentaires sur les sorties des analyses multidimensionnelles. Par exemple, Tayrac *et al.* (2009), dans l'étude de l'ACP d'un tableau de type sujets$\times$gènes, ont proposé une stratégie de représentation de clusters de gènes, basée sur l'utilisation de clusters de gènes Gene Ontology. Dans cette stratégie chaque terme GO est utilisé de façon indépendante pour constituer un cluster. Un cluster s'interprète donc directement et correspond à un terme GO dans ce cas. La grille d'analyse proposée par Tayrac *et al.* (2009) est intéressante dans la mesure où elle permet de synthétiser l'information au

niveau des gènes au moyen des termes GO, néanmoins elle peut s'avérer insuffisante et les clusters obtenus sont d'interprétation parfois délicate.

Nous cherchons donc à intégrer de façon active l'information biologique. Nous proposons au préalable de présenter le principe de la méthodologie que nous cherchons à développer.

3.2 Principe

La coexpression entre deux gènes peut avoir une multiplicité d'interprétation. Pour simplifier, nous proposons de considérer un nouveau schéma de relation entre coexpression et liaison biologique. Pour schématiser, disons que la coexpression de deux gènes peut résulter de deux phénomènes, soit d'un véritable lien biologique (du type réseau de régulation génique), soit de l'activation parallèle et indépendante de différentes réponses biologiques à une même condition expérimentale. Nous souhaitons différencier, au moyen d'une information biologique extérieure sur les gènes, la coexpression résultant d'une liaison biologique, de la coexpression résultant de mécanismes parallèles. Ainsi, nous considérons que si deux gènes coexprimés sont associés à un même ensemble de termes GO, cela traduit l'existence d'un véritable lien biologique entre ces deux gènes. Au contraire, s'ils ne partagent aucun terme GO, cela traduit l'implication de ces deux gènes dans des mécanismes biologiques parallèles.

Pour intégrer de l'information biologique dans les données d'expression, il est nécessaire de trouver un codage de l'information biologique pour pouvoir l'utiliser dans un contexte statistique. Nous avons choisi de coder les associations entre gènes et termes GO dans une matrice, nommée $\mathbf{T}$ qui croise les gènes en lignes et les termes GO en colonnes. A l'intersection d'une ligne et d'une colonne, se trouve un 1 si le gène est associé au terme, 0 sinon. Cela nous permet ainsi de définir le profil ou la signature fonctionnelle d'un gène comme l'ensemble des fonctions biologiques représentées par les termes GO auxquels le gène est associé, cela correspond à une ligne de $\mathbf{T}$. Par ailleurs, nous considérons le profil d'expression des gènes qui correspond à une ligne du tableau gènes$\times$sujets. En effet comme nous avons établi section 1.2.2 du chapitre 2, nous allons utiliser cette seconde orientation du tableau de données qui présente un intérêt technique. Ainsi un gène est caractérisé par deux signatures, une signature fonctionnelle et une signature d'expression.

Notre objectif est d'imaginer une distance entre gènes qui permettrait de quantifier la similarité de leurs signatures fonctionnelles, tout en tenant compte du fait qu'ils soient coexprimés ou non. Ainsi, nous cherchons à obtenir des clusters qui rassemblent des gènes à la fois coexprimés et biologiquement cohérents. Techniquement, cela peut se traduire par une combinaison des signatures fonctionnelle et d'expression des gènes pour en faire une unique signature qui permettrait de définir une nouvelle distance entre gènes qui serait utilisée dans une perspective de clustering. La prise en compte de ces deux sources d'information peut se faire soit de façon symétrique en équilibrant le rôle des deux signatures au moyen d'une analyse factorielle multiple (Escofier et Pagès, 2008), soit de

façon dissymétrique à travers l'analyse canonique des correspondances (Ter Braak, 1986) en privilégiant la coexpression.

3.3 Approche symétrique : l'AFM

Avec une approche symétrique, nous cherchons à combiner les signatures fonctionnelle et d'expression en une unique signature qui permettrait d'obtenir des clusters de gènes, à la fois coexprimés et cohérents d'un point de vue de l'information biologique. Nous cherchons à définir une distance combinant les deux signatures au moyen d'une analyse factorielle multiple (AFM) sur les tableaux gènes$\times$sujets et $\mathbf{T}$. Dans cette analyse, le poids des deux tableaux est équilibré. Nous appliquons ensuite un algorithme de clustering sur l'ensemble des coordonnées factorielles, ce qui conduit à l'obtention de clusters de gènes basés sur cette distance. Nous nous intéressons enfin aux caractéristiques des clusters obtenus sur la base de cette distance.

Nous observons un premier résultat très étonnant, une grande partie des clusters obtenus rassemble des gènes qui ne sont pas coexprimés. Nous pouvons donc nous demander si cette analyse symétrique qui vise à donner le même poids aux données d'expression et à l'information biologique, remplit bien cette fonction. Nous proposons, pour comprendre ce résultat, de constituer des clusters de gènes en ne se basant que sur l'information biologique. Pour cela, nous réalisons une analyse factorielle des correspondances (AFC) du tableau $\mathbf{T}$ et appliquons un algorithme de clustering sur les coordonnées factorielles des gènes. Si nous comparons la partition des gènes obtenue par AFM et la partition basée sur la seule information biologique obtenue par AFC, ces deux partitions sont presque rigoureusement les mêmes.

Ces résultats étonnants peuvent néanmoins s'expliquer. Le fait que les partitions soient identiques résulte directement de la similarité des distances entre gènes dans les espaces issus de l'AFC de $\mathbf{T}$ et de l'AFM. Pour comprendre ce phénomène, nous pouvons nous centrer sur l'interprétation de l'AFM. Pour interpréter l'AFM sur les deux tableaux $\mathbf{T}$ et gènes$\times$sujets, nous pouvons regarder la valeur de la première valeur propre de l'analyse qui varie entre 1, s'il y a peu de structure commune entre les deux tableaux, et 2, s'il y a une très forte structure commune entre les deux tableaux. Effectivement, la première valeur propre de l'analyse étant de 1.13, cela signifie qu'il y a peu de structure commune entre les tableaux $\mathbf{T}$ et gènes$\times$sujets. De plus, nous pouvons également comparer les axes des analyses séparées aux axes de l'AFM, les axes des analyses séparées étant les axes de l'ACP du tableau gènes$\times$sujets et les axes de l'AFC du tableau $\mathbf{T}$. Hormis le premier axe de l'AFM, la corrélation maximum entre un axe de l'ACP et un axe de l'AFM est de 0.2. Au contraire, beaucoup d'axes de l'AFM sont fortement corrélés à un des axes de l'AFC. A titre indicatif, 22 des 30 premiers axes de l'AFM présentent une corrélation supérieure à 0.9 avec un axe de l'AFC. En conséquence, seul le premier axe de l'AFM est commun aux deux groupes de variables, tandis que les autres axes sont plus spécifiques de l'AFC. Cela résulte directement de la pondération de l'AFM. Nous nous intéressons aux valeurs propres des deux analyses séparées, mais pondérées par les premières valeurs

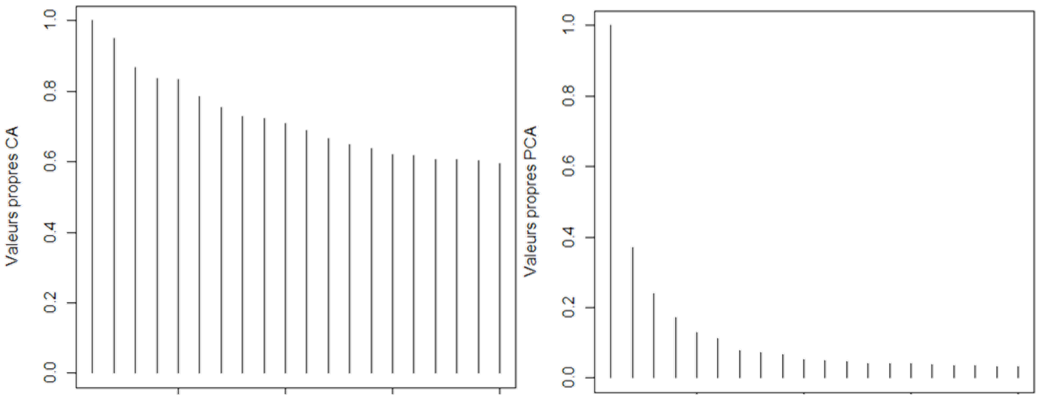


FIGURE 3 – Éboulis des valeurs propres de l'AFC et de l'ACP pondérées par leur première valeur propre respective.

propres respectives (figure 3). Tous les axes de l'AFC portent plus d'inertie que le second axe de l'ACP. Ainsi, les axes de l'AFC jouent un rôle prépondérant dans la construction des axes de l'AFM qui sont finalement plus spécifiques de l'AFC.

Cette stratégie de combinaison symétrique des signatures d'expression et fonctionnelle n'est pas satisfaisante parce qu'en donnant le même poids (en termes d'inertie) aux deux signatures, la combinaison symétrique avantage la signature fonctionnelle qui est plus multidimensionnelle. Nous proposons à présent une autre stratégie qui consiste à combiner les signatures d'expression et fonctionnelle de façon dissymétrique, en donnant plus de poids à la signature d'expression.

3.4 Approche dissymétrique : l'ACC

Nous cherchons à combiner les signatures d'expression et fonctionnelle en une unique signature, mais en donnant plus de poids à la signature d'expression. Nous définissons ainsi une distance entre gènes qui garantit avant tout la coexpression des gènes mais qui est également basée sur les ressemblances entre signatures fonctionnelles. Nous proposons pour cela de réaliser une analyse canonique des correspondances (ACC) combinant les tableaux gènes$\times$sujets et $\mathbf{T}$. Expliquons succinctement le principe de cette analyse. Nous considérons les gènes dans l'espace formé par les sujets dans l'ACP du tableau gènes$\times$sujets : cet espace est celui de référence. Puis les fonctions biologiques (colonnes du tableau $\mathbf{T}$) sont projetées dans cet espace comme une moyenne pondérée des gènes qui lui sont associés. Ensuite, nous recherchons les plus grandes directions d'inertie au sein de ces fonctions. Enfin, les gènes sont projetés sur ces directions. Une fois cette distance définie, nous obtenons des clusters de gènes en appliquant un algorithme de clustering sur les coordonnées des gènes ainsi projetés sur les axes de l'ACC.

Intéressons nous aux caractéristiques des clusters de gènes obtenus. Nous nous attendons à obtenir des clusters rassemblant des gènes avant tout coexprimés et présentant des signatures fonctionnelles assez proches. Effectivement, les clusters

obtenus rassemblent des gènes coexprimés. L'espace de référence de l'ACC est bien celui de l'ACP, ce qui permet de garantir la coexpression. Cependant, les gènes de ces clusters ne présentent pas de signatures fonctionnelles très proches.

Les deux tableaux ayant peu de structure commune, du fait de la très grande multidimensionnalité de l'information biologique, il est très difficile de combiner les signatures d'expression et fonctionnelle en une unique signature, au moyen d'une analyse multidimensionnelle. C'est pourquoi nous nous sommes tournés vers une solution algorithmique, plus empirique, mais qui donne des résultats prometteurs. Cette solution est proposée dans l'article suivant :
Verbanck, M., Lê, S., & Pagès, J. (2013). A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. *BMC Bioinformatics*, **14**, 42. (Highly Accessed)

A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data (Verbanck *et al.*, 2013b)

Résumé : Les stratégies classiques de clustering sont uniquement basées sur l'utilisation des données d'expression, directement comme dans les Heatmaps ou indirectement comme dans le clustering basé sur des réseaux de coexpression par exemple. Cependant, les stratégies classiques ne sont peut-être pas suffisantes pour extraire toutes les relations potentielles entre gènes. Nous proposons un nouvel algorithme de classification non supervisée de gènes basée sur l'intégration d'information biologique extérieure, telle que des annotations Gene Ontology, dans les données d'expression. Nous introduisons une nouvelle distance entre gènes, deux gènes sont proches s'ils présentent à la fois des signatures d'expression et des signatures fonctionnelles similaires. Ensuite, un algorithme (par exemple K-means) est utilisé pour obtenir des clusters de gènes. De plus, nous proposons une procédure d'évaluation automatique des clusters de gènes. Cette procédure est basée sur deux indicateurs qui mesurent la coexpression globale et l'homogénéité biologique des clusters de gènes. Ils sont associés à des tests d'hypothèse qui permettent de compléter les indicateurs avec une probabilité critique. Notre algorithme de clustering est comparé au clustering Heatmap et au clustering basé sur un réseau de coexpression, à la fois sur des données simulées et réelles. Dans les deux cas, notre algorithme est plus performant que les autres méthodologies puisqu'il fournit la plus grande proportion de clusters de gènes significativement coexprimés et biologiquement homogènes, qui sont de bons candidats à l'interprétation. Ainsi, nous espérons que l'interprétation de ces clusters pourra aider les biologistes à formuler de nouvelles hypothèses sur les relations entre gènes.

Verbanck *et al. BMC Bioinformatics* 2013, **14**:42
http://www.biomedcentral.com/1471-2105/14/42

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data

Marie Verbanck*, Sébastien Lê and Jérôme Pagès

Abstract

Background: Gene clustering algorithms are massively used by biologists when analysing omics data. Classical gene clustering strategies are based on the use of expression data only, directly as in Heatmaps, or indirectly as in clustering based on coexpression networks for instance. However, the classical strategies may not be sufficient to bring out all potential relationships amongst genes.

Results: We propose a new unsupervised gene clustering algorithm based on the integration of external biological knowledge, such as Gene Ontology annotations, into expression data. We introduce a new distance between genes which consists in integrating biological knowledge into the analysis of expression data. Therefore, two genes are close if they have both similar expression profiles and similar functional profiles at once. Then a classical algorithm (e.g. K-means) is used to obtain gene clusters. In addition, we propose an automatic evaluation procedure of gene clusters. This procedure is based on two indicators which measure the global coexpression and biological homogeneity of gene clusters. They are associated with hypothesis testing which allows to complement each indicator with a p-value. Our clustering algorithm is compared to the Heatmap clustering and the clustering based on gene coexpression network, both on simulated and real data. In both cases, it outperforms the other methodologies as it provides the highest proportion of significantly coexpressed and biologically homogeneous gene clusters, which are good candidates for interpretation.

Conclusion: Our new clustering algorithm provides a higher proportion of good candidates for interpretation. Therefore, we expect the interpretation of these clusters to help biologists to formulate new hypothesis on the relationships amongst genes.

Background

Since omics data such as transcriptome profiling data provide measures about a considerable number of genes, data are classically decomposed to a more comprehensible level by clustering genes into modules. Among the unsupervised clustering strategies we can recall the two techniques that are principally used: Heatmaps [1] which consist in hierarchical classification on both subjects and gene expressions, and clustering based on coexpression networks [2]. Gene clustering is not only practical since it reduces the number of objects to study, but is also expected to convey a certain biological reality. In fact, we expect the similarities between gene expressions to reflect similarity between gene functions. Gene clusters are then interpreted in order to generate new hypotheses about the functional roles of genes and their relationships.

In practice, to interpret gene clusters, external biological knowledge such as Gene Ontology (GO) information [3] is used. The most classical procedure consists of gene set enrichment analysis with the aim to characterise each cluster by a set of biological functions. Attempts to improve gene set enrichment analysis have been proposed, for instance Bauer et al. [4] proposed a Bayesian enrichment analysis. The latter consists in representing GO terms into a Bayesian network and the response of each gene, in terms of expression, is modelled as a function of the activation of GO terms. In Multivariate Analysis (MVA), some attempts to directly superimpose

*Correspondence: marie.verbanck@agrocampus-ouest.fr
Applied Mathematics Department, Agrocampus Ouest, 65, rue de Saint-Brieuc, Rennes, France

biological knowledge on the outputs of MVA exist [5,6]. The objective is to facilitate the interpretation of gene expressions, or gene clusters, as MVA provides distance matrices that can be used for clustering.

In these methodologies, gene clusters are obtained on the basis of expression data only and biological knowledge is a posteriori used to make the most of the clusters. The limits of such procedures are clear: clustering genes on the basis of expression data only allows to isolate coexpressed, however not necessarily biologically coherent units [7,8]. Indeed, a clustering structure can only be as good as the distance/similarity matrix it is based on. Hence, the idea of actively integrating biological knowledge into expression data, to isolate more meaningful biological entities.

In other contexts, this issue of actively integrating biological knowledge into expression data has been covered. In the purpose of biological networks inference, Kashima et al. [9] proposed a semi-supervised learning method. The similarity between expression profiles and amino acid sequences in a given species is reinforced if the same similarity is observed amongst a cousin species. In order to predict gene functional classes, such as the associations between genes and GO terms, Azuaje et al. [10] combine two types of information: gene expression profile similarity and a GO-based similarity. The average of both similarity indexes is used to cluster genes. With the same objective of predicting gene functional classes, in Li et al. [11], expression data are combined with biological knowledge by considering subsets of genes associated with one same functional annotation. The subsets of genes are then clustered on the basis of their expression profile similarities.

The objective of the paper is to propose a new unsupervised clustering algorithm based on a new distance between genes that actively integrates external biological knowledge into expression data. A cluster is considered as satisfying if it gathers coexpressed genes that are implicated into similar biological functions according to the biological knowledge. Such a cluster is expected to be biologically interesting and becomes a good candidate for biological interpretation.

In practice, we introduce the notion of coexpressed biological functions which allows the integration of an information of coexpression within the functional annotations. Combining expression data with GO annotations defines a new distance between genes. Two genes are close if they are coexpressed and implicated into the same set of biological functions at once. Afterwards a classical clustering algorithm (K-means or hierarchical ascending classification) is used to obtain gene clusters. In this paper we will emphasize the biological principle supporting the methodology and discuss the distance we propose.

To complement the clustering procedure, we propose an automatic validation procedure of gene clusters to facilitate their interpretation. The aim of this procedure is to highlight good candidates for interpretation which are clusters of significantly coexpressed and significantly biologically related genes. It is based on two indicators associated with hypothesis testing. One indicator measures the coexpression of the genes within a cluster, whereas the other quantifies its biological homogeneity.

The R code which is used to perform all analyses is available in the form of an R package at http://marie.verbanck.free.fr/packages/.

Method

Integration of biological knowledge into expression data: biological principle

Let us recall that most of the classical gene clustering strategies are based on expression data only. Expression data may be used directly as in Heatmaps, or indirectly in the case of clustering based on coexpression networks. Clusters thus obtained are candidates for interpretation and remain to be biologically characterised. The biological characterisation is done using external biological knowledge, such as Gene Ontology annotations. These are established according to experiments reported in the literature, or deduced by Bioinformatics. This classical approach relies on two implicit hypotheses. Firstly, the biological characterisation of coexpressed clusters implicates that biological connections systematically exist between coexpressed genes. Secondly, the biological characterisation is purely based on external biological knowledge, therefore, part of the external biological knowledge is expected to be related to the experiment in the study.

The first hypothesis may be questionable [7,8] and in this paper we consider a new point of view on the link between coexpression and biological connections. Broadly speaking, coexpression between two genes may result from two phenomena, either a genuine biological connection (e.g. from a true gene regulation network), or the parallel and independent activation of different biological responses to the same experimental condition. To differentiate those two situations, we propose to give more credit to the second hypothesis and then to actively rely on external biological knowledge. Therefore, we consider that if two coexpressed genes have already been characterised as biologically related in the existing biological knowledge, their coexpression is more likely to reflect a genuine biological connection.

In practice, we use the ontology related to "Biological Process" of GO annotations which provides for each gene a list of biological functions which the gene is involved in: henceforth, this list will be called *functional profile* of the gene. Therefore, if two coexpressed genes are associated with similar functional profiles, their coexpression is presumed to result from a genuine biological connection. On the contrary, if two coexpressed genes

have totally divergent functional profiles, their coexpression may result from the parallel activation of different biological responses.

Unsupervised gene clustering algorithm

In this section, we propose a new distance between genes, that fits the exposed biological principle, and to be used in a clustering perspective. This distance allows to quantify both the coexpression and the similarity of functional profiles between two genes.

Encoding of the biological knowledge

Let us consider K genes and J GO terms. The associations between genes and GO annotations are encoded in a binary matrix $T \in \mathcal{M}(K,J)$, where each line k represents one of the K genes and each column j one of the J GO terms: the general term T_{kj} equals 1 if the gene k is associated with the GO term j and 0 else wise (Figure 1). A row k of the matrix can be interpreted as a gene functional profile which is the set of biological functions the gene is associated with. A column j of the matrix represents a biological function that can be assimilated to the subset of genes that are associated with the function in question. Let $K^j = \{k | T_{kj} = 1\}$ be the subset of genes that are associated with the function j.

A new distance between genes: coexpressed biological functions

In order to fit the previously exposed biological principle, we define a distance that quantifies the similarity of functional profiles $\{T_{kj}; j \in J\}$ of coexpressed genes. To do so, we apply a constraint on the biological knowledge by defining a *coexpressed biological function* as the restriction of the function to the only genes that are coexpressed. In other words, if K^j can be split up into L_j coexpressed clusters, that will lead to as many coexpressed biological functions to be considered. In order to obtain these coexpressed biological functions, we propose the following algorithm based on hierarchical clustering.

For each biological function j:

1. a distance matrix between the genes of K^j based on Pearson's correlation coefficient is computed. The distance between two genes k and k' may be expressed as follows:

$$d_G(k,k') = 1 - \frac{1}{I}\sum_{i=1}^{I}\left(\frac{G_{ik} - G_{.k}}{S_k}\right)\left(\frac{G_{ik'} - G_{.k'}}{S_{k'}}\right) \quad (1)$$

   where I is the number of samples, G_{ik} and $G_{ik'}$ are respectively the expression of genes k and k' for sample i, $G_{.k}$ and $G_{.k'}$ are respectively the mean of the I expression values of genes k and k', S_k and $S_{k'}$ are respectively the standard deviation of the I expression values of genes k and k'.
2. A hierarchical clustering procedure is performed on the previously defined distance matrix (1): let $P^j = \{K_1^j; ...; K_l^j; ...; K_{L_j}^j\}$ be a partition on K^j in L_j clusters. For all $l = 1, \ldots, L_j$, K_l^j is comprised of coexpressed genes.
3. We build a matrix $T^j \in \mathcal{M}(K, L_j)$ by splitting up the j^{th} column of T into L_j columns. In T^j each line k represents one of the K genes and each column is a dummy variable such as T_{kl}^j equals 1 if the gene k belongs to K_l^j and 0 else wise: a column of T^j can be interpreted as a coexpressed biological function.

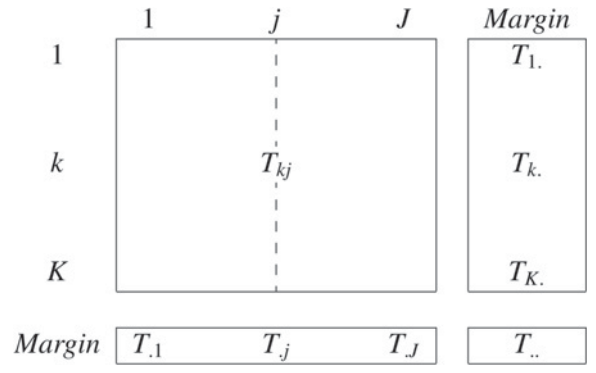


Figure 1 Matrix *T*: coding the associations between genes and biological functions. The associations between genes and biological functions are synthesised in the matrix *T*. Each row represents a gene functional profile, whereas each column represents the associations between a biological function and genes. The general term T_{kj} equals 1 if the gene k is associated with the biological function j, 0 else wise. The row margin $T_{k.}$ is the number of biological functions the gene k is associated with. The column margin $T_{.j}$ is the number of genes the function j is associated with. Finally, $T_{..}$ is equal to the total number of associations between genes and biological functions.

We define T_{coexp} as the juxtaposition of all J matrices T^j (Figure 2). T_{coexp} results from combining both types of information. The analysis of T_{coexp} allows to study the degree of similarity of gene functional profiles under the condition of coexpression. Therefore a new distance between genes can be calculated from T_{coexp}:

$$d_{T_{coexp}}(k,k') = \sum_{j=1}^{J}\sum_{l=1}^{L_j}\frac{T_{..}}{card(K_l^j)}\left(\frac{T_{kj}}{T_{k.}}\mathbb{1}_{k\in K_l^j} - \frac{T_{k'j}}{T_{k'.}}\mathbb{1}_{k'\in K_l^j}\right)^2 \quad (2)$$

where $T_{k.}$ and $T_{k'.}$ are respectively the row margins associated with the genes k and k', $T_{..}$ is the total number of associations between genes and biological functions and $\mathbb{1}_{k\in K_l^j}$ a dummy variable which equals 1 if $k \in K_l^j$, 0 else wise. The genes k and k' are both associated with

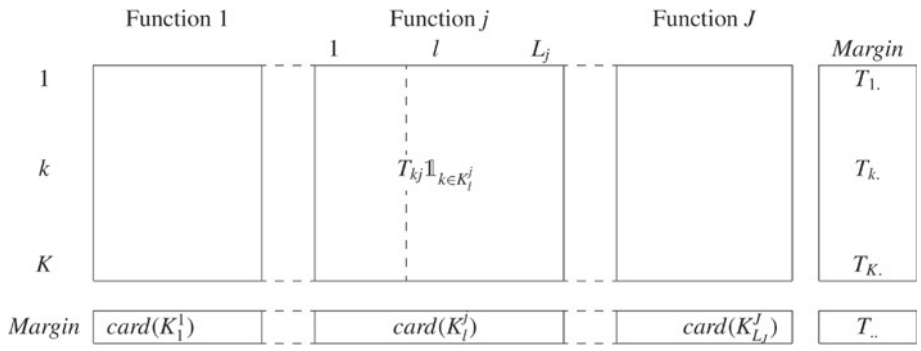


Figure 2 Matrix T_{coexp}: decomposition of the matrix *T*. Decomposing biological functions into coexpressed biological functions leads to build the matrix T_{coexp} where a row represents a gene and a column a coexpressed biological function. The general term of T_{coexp}, $T_{kj}\mathbb{1}_{k\in K_l^j}$ equals 1 if the gene *k* is associated with the function *j* and if it belongs to the cluster K_l^j, 0 else wise. The column margin of the coexpressed biological function *l* is equal to the number of genes in the corresponding cluster, that is $card(K_l^j)$. In addition, for every function *j*, the sum of the column margins associated with the coexpressed biological functions derived from *j* is equal to the column margin associated with the function *j*: $\sum_{l=1}^{L_j} card(K_l^j) = T_j$. Finally, we can remark that the row margins and the total number of associations are equal to those from *T*.

j: if they are not coexpressed they do not belong to the same coexpressed cluster of P^j. In this case, the j^{th} term of the distance calculation (2) is high. Thus, genes which have similar expression profiles and similar functional profiles are close. This distance corresponds to the distance between genes in the Correspondence Analysis of T_{coexp}.

Technical note 1: in step 2, P^j is the partition in L_j coexpressed clusters of the genes associated with the biological function j. P^j is determined by cutting the classification tree. Cutting the classification tree provides a partition and allows to calculate the sum of the intra-cluster inertias for the partition in question. The relative loss of inertia is calculated between the partition in L clusters and the partition in $L+1$ clusters as $\frac{\sum_{l=1}^{L+1} inertia(l)}{\sum_{l=1}^{L} inertia(l)}$. P^j is obtained by cutting the classification tree to obtain the partition with the higher relative loss of inertia.

Technical note 2: in the particular case where all genes associated with j, are coexpressed, j is then considered as a coexpressed biological function. We add a step 0. consisting in filtering biological functions: it allows to define whether a biological function j can be considered as coexpressed. For that matter, the coexpression of the subset of genes associated with j is tested by calculating the p-value of the coexpression indicator according to the procedure presented in the following section. If this p-value is lower than a chosen threshold (e.g. 10%), the function in question is considered as a coexpressed function and will not be split up in T_{coexp}, but conserved as it is.

Note: in a totally different context, with the aim of predicting gene functional classes, Li et al. [11] proposed a fuzzy near-cluster algorithm base on the idea of detecting homogeneous co-expressed gene subgroups in heterogeneous functional class which is close to ours. This detection allows them to have a better prediction of gene functional classes.

Obtaining gene clusters

To obtain gene clusters, a clustering algorithm, such as K-means or hierarchical ascending classification, is then applied to the distance matrix. We expect, from this procedure, to obtain clusters of coexpressed and biologically related genes.

Evaluation of gene clusters

For a cluster to be a good candidate for interpretation, it has to gather coexpressed and biologically related genes. Classical evaluation procedures focus on what can be called the *biological homogeneity* of a cluster and its characterisation by biological functions. However, in our clustering procedure, coexpression is necessarily competing with biological homogeneity, as both types of information are actively combined. Therefore, we propose an evaluation procedure of gene clusters based on two indicators: a coexpression and a biological homogeneity indicator associated with hypothesis testing.

Coexpression indicator

Coexpression is defined as a positive correlation between two genes. Indeed, if two genes are positively correlated, they are over- and under-expressed in the same experimental conditions. We want to find a coexpression indicator (CI) that synthesizes correlations within a cluster. We consider an empirical, but convenient, indicator which is

the average of correlations between the genes of the same cluster K_l. This indicator is calculated as follows:

$$CI(K_l) = \frac{1}{\frac{card(K_l)(card(K_l)-1)}{2}} \sum_{k|k\in K_l} \times \left(\sum_{k'|k'\in K_l, k'>k} \frac{1}{I} \sum_{i=1}^{I} \left(\frac{G_{ik}-G_{.k}}{S_k}\right)\left(\frac{G_{ik'}-G_{.k'}}{S_{k'}}\right)\right) \quad (3)$$

where I is the number of samples, G_{ik} and $G_{ik'}$ are respectively the expression for the sample i of the genes k and k', $G_{.k}$ and $G_{.k'}$ are respectively the mean of the I expression values of the genes k and k', S_k and $S_{k'}$ are respectively the standard deviation of the I expression values of the genes k and k'.

The coexpression indicator indeed offers a measure of the global situation of coexpression of gene clusters. It ranges from $-\frac{1}{3}$ to 1 (See Appendix 1). If all genes are perfectly coexpressed, the indicator equals 1. On the contrary, let us considered a cluster whose genes are not coexpressed, to such an extent that two sub-clusters are distinguished: within each sub-cluster, genes are positively correlated, and between sub-clusters, they are negatively correlated. In this case, the indicator is close to 0 and might be less than 0.

Biological homogeneity indicator

We aim at defining a biological homogeneity indicator based on the similarity of gene functional profiles. Classically, the biological homogeneity of a gene cluster is appraised by the number and the nature of enriched biological functions which are associated with it. However, the characterisation of a cluster by enrichment tests does not guarantee the similarity of functional profiles as enrichment tests are conducted separately for each biological function. Datta & Datta [12] proposed a multidimensional biological homogeneity indicator with the objective to evaluate the whole clustering procedure, not the clusters themselves. We adapt this idea to measure the biological homogeneity of gene clusters. We consider as the biological homogeneity indicator (BHI) a coefficient derived from Cramér's V coefficient [13] which offers a measure of the degree of similarity of functional profiles of genes from K_l. This indicator is calculated as:

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(\sum_{j=1}^{J} \frac{\left(T_{kj} - \frac{T_{k.}T_{.j}}{T_{..}}\right)^2}{\frac{T_{k.}T_{.j}}{T_{..}}}\right)}{T_{..}(card(K_l)-1)}} \quad (4)$$

where T_{kj} equals 1 if the gene k is associated with the biological function j, 0 else wise, $T_{k.}$ is the row margin associated with the gene k.

The biological homogeneity indicator varies between 0 and 1 (See Appendix 2). Therefore, if all genes from a cluster have perfectly similar functional profiles, the biological homogeneity indicator equals 1. On the contrary, if none of the genes have similar functional profiles to such an extent that none of the biological functions is associated with two of the genes from K_l, then the biological homogeneity indicator equals 0.

Although this indicator has its limits, as biological homogeneity should principally rely on biological interpretation, nevertheless, it happens to be useful to automatically be able to assess the biological interest of gene clusters.

Hypothesis testing procedure

We complement the indicators with a hypothesis testing procedure, which is all the more legitimate as both indicators strongly depend on the size of the cluster:

- coexpression indicator: in its calculation (3) a division by $\frac{card(K_l)(card(K_l)-1)}{2}$ is performed, CI's value mechanically decreases with the size of clusters
- biological homogeneity indicator: a division by $card(K_l)-1$ is performed in the second term of its calculation (4), and as this second term varies between 0 and 1, BHI's value mechanically increases with the size of the cluster

The objective is to evaluate to what extent a methodology provides clusters whose coexpression and biological homogeneity are higher than in a situation of random clustering. Consequently, random clustering corresponds to the null hypothesis of the test, and the values of the indicators of random clusters are taken as a reference situation. In practice, to associate a p-value to the cluster K_l for one indicator, clusters of the same size are constituted by simply drawing genes without replacement. The indicator is then calculated for each cluster and a distribution of the values of the indicator under the null hypothesis is thus obtained. As usual, the observed value, corresponding to the value of the indicator for the cluster to be tested, is positioned in the corresponding distribution under the null hypothesis. Ultimately, the p-value is estimated by the proportion of randomly constituted clusters whose indicator value is superior to the observed value.

Note 1: the interest of the procedure resides in the way distributions under the null hypothesis are obtained. As the calculation of the indicators remains based on real data, the distributions under the null hypothesis respect the distributions of the data.

Note 2: obviously clusters composed of one single gene are not tested.

Results

As we propose a new unsupervised clustering algorithm associated with an automatic evaluation of the clusters, we validate the whole methodology on simulated, and real data sets, by comparing it with two of the most classically used gene clustering strategies. On the one hand we compare it with clusters stemming from a Heatmap of the expression data. On the other hand, we choose to generate a coexpression network from the expression data using Weighted Gene Coexpression Network (WGCNA) [2]. The coexpression network allows to calculate a dissimilarity matrix between genes based on the topological overlap of the nodes of the network. Finally a hierarchical clustering algorithm is computed on the dissimilarity matrix and provides gene clusters.

Simulation study

Simulated data sets

In this section, we explain how to simulate expression and GO data sets.

To simulate expression data, we use the same procedure as in [14]. An expression data matrix G_{sim}, constituted of K genes and I samples, is simulated from random drawing in a multivariate Gaussian distribution with a certain correlation structure so that we have underlying clusters of coexpressed genes. Since this way of simulating numerical data is quite classical, we rather insist on the simulation of GO annotation data which is not common in the literature.

To simulate GO annotation data we fit the biological principle previously exposed: GO annotations are constituted by information that can be related to the experiment in the study and information that is not. In other words, one part of the simulated GO annotations must have a structure which is similar to the structure of the expression data, and the other must have a random structure. Thus, a simulated GO matrix T_{sim} is obtained by juxtaposing two types of matrices:

- T^e_{sim}: its gene functional profiles emulate gene expression profiles, thus when two genes have similar expression profiles in G_{sim}, they have similar functional profiles in T^e_{sim}
- T^r_{sim}: its genes functional profiles are not related to gene expression profiles

In practice, to obtain T^e_{sim}, first we build a gene classification tree based on correlations between their expression profiles only. Then we consider each node j of the classification tree as a biological function. If the gene k is associated with the node j of the classification tree, $T^e_{sim}(k, j) = 1$, 0 else wise. As a result, genes that have similar expression profiles mechanically share close functional profiles. To obtain T^r_{sim}, we juxtapose r times the matrix T^e_{sim} and independently permute rows within each column, where r is an integer representing the intensity of randomness of T_{sim}: concretely, there are r times more random biological functions than structured biological functions in T_{sim}.

This way of generating the similar matrix of T^e_{sim} is chosen as it mimics the hierarchical structure of GO information. This way of generating the random matrix T^r_{sim} allows to conserve the margins of biological functions, what is important as these margins represent the number of genes that are associated with the functions and may be interpreted as a degree of specificity of the functions.

Results

In practice, we apply the three methods to simulated data sets. We consider two sizes of simulated expression data. A first type composed of 10 individuals and 300 genes for which we obtain a partition in 20 clusters for each method. A second type composed of 25 individuals and 1000 genes for which we obtain a partition in 100 clusters for each method. With both types of simulated expression data sets, we associate simulated GO annotations whose intensity of randomness ranges from 1 to 3. For each configuration 100 data sets are generated.

Whatever the clustering method, we associate with each cluster, two p-values corresponding each to the coexpression indicator and the biological homogeneity indicator. For a given partition, we measure the proportion of clusters which are:

- significantly coexpressed: p-value associated with the CI lower than a chosen threshold
- significantly biologically homogeneous: p-value associated with the BHI lower than a chosen threshold
- both significantly coexpressed and biologically homogeneous: both p-values associated with the CI and the BHI lower than a chosen threshold

Results are gathered in Table 1. On average, all three methods provide partitions with a high proportion of significantly coexpressed clusters. This proportion does not depend on the intensity of randomness for Heatmap and WGCNA. However, for our clustering algorithm, we observe a slight decrease in the proportion of significantly coexpressed clusters when the intensity of randomness increases. This is expected as coexpression is competing even more with biological homogeneity when the intensity of randomness is high.

On average, partitions stemming from Heatmaps have low proportions of clusters which are significantly biologically homogeneous. This proportion severely decreases when the intensity of randomness increases. Taking into account a network structure behind gene expressions is beneficial since it provides a greater proportion

Table 1 Results of the simulation study

			Coexpression indicator			Biological homogeneity indicator			Both		
I	*K*	*r*	**Heatmap**	**WGCNA**	**Integration**	**Heatmap**	**WGCNA**	**Integration**	**Heatmap**	**WGCNA**	**Integration**
10	300	1	92.15	94.90	98.65	65.50	81.5	89.5	64.60	78.95	88.80
10	300	2	92.31	94.80	96.55	50.40	60.15	67.25	49.75	58.30	66.25
10	300	3	92.00	95.32	94.52	36.77	45.81	54.03	36.61	45.00	53.39
25	1000	1	88.70	99.12	91.33	7.67	28.00	45.44	7.35	27.09	44.72
25	1000	2	90.25	99.12	90.55	3.79	11.89	29.62	3.54	11.17	28.95
25	1000	3	89.00	98.99	85.67	1.94	3.55	18.66	1.80	3.34	18.06

Results of the simulation study for the three clustering algorithms: Heatmap classification (Heatmap), clustering based on coexpression network (WGCNA) and our clustering algorithm (Integration). The simulated data sets vary according to the number of samples (*I*), the number of genes (*K*) and the intensity of randomness (*r*). We give the average proportion of clusters (%), among a given partition, which are significantly coexpressed (CI), biologically homogeneous (BHI) or both coexpressed and biologically homogeneous (Both). Let us take the example of simulated expression data sets with 10 individuals and 300 variables, associated with simulated GO annotations with an intensity of randomness of 1. On average the Heatmaps of these data sets provide partitions with 92.15% of significantly coexpressed clusters.

of significantly biologically homogeneous clusters than Heatmap. However, the proportion of biologically homogeneous clusters provided by WGCNA literally drops when the intensity of randomness is very high. Our clustering algorithm provides a reasonably high proportion of biologically homogeneous clusters even when the intensity of randomness equals 3.

If we focus on the proportion of clusters which are both significantly coexpressed and biologically homogeneous, our clustering algorithm outperforms the other two methods.

Analysis of the chicken data set

The methodology is applied to an example of transcriptomic data set which is related to a published data set [15]. The aim, through this experiment, is to understand the genetic mechanisms implemented in reply to fasting in chickens. Therefore, the expression of about 12 000 hepatic genes was collected in 27 chickens submitted to 4 nutritional statuses: 16-hour fasting "F16", 16-hour fasting followed by a 5-hour renutrition phase "F16R5", 16-hour fasting followed by a 16-hour renutrition phase "F16R16" and finally, a continuously fed status "F". We choose in our example to perform a selection of genes whose expression varies according to the experimental factor, which led us to retain about 3600 genes thanks to the Factor Analysis for Multiple Testing method [16].

In addition, similarly to Busold et al. [5], we use GO information where the hierarchical structure amongst GO terms is taken into account: when a gene is associated with a term, it is automatically associated with its parents.

As in the simulation study, we perform three gene clusterings corresponding to a Heatmap, a clustering based on a coexpression network (WGCNA) and our own clustering procedure. We choose to set the number of clusters obtained from each procedure to 200. For a given partition, we associate with each cluster two p-values for the coexpression indicator and the biological homogeneity indicator which are visualised in a joint graph. In Figure 3, a point represents a cluster whose value on the x-axis is equal to the coexpression indicator p-value and whose value on the y-axis is equal to the biological homogeneity indicator p-value. In addition, Table 2 provides the proportion of clusters, amongst each one of the three partitions, which are significantly coexpressed (CI), biologically homogeneous (BHI) or both coexpressed and biologically homogeneous (Both), as in the simulation study.

Firstly, the partition provided by the Heatmap is constituted of a large majority of clusters which are significantly coexpressed (91.50%). However a small proportion of the clusters are significantly biologically homogeneous to such an extent that p-values associated with the BHI seem to be distributed according to a uniform distribution. A QQ-plot (Figure 4) actually confirms that the p-value distribution associated with the biological homogeneity indicator can be considered as uniform, which corresponds to a distribution followed by p-values under the null hypothesis. Therefore, Heatmap clustering may come down to cluster genes independently from any biological homogeneity.

Secondly, compared to Heatmap, considering a coexpression network considerably improves the results. Thus WGCNA provides a much higher proportion of biologically homogeneous clusters (68%). However, the proportion of coexpressed clusters decreases. Ultimately WGCNA provides a reasonable proportion of good candidates for interpretation (46%).

Thirdly, with our own clustering algorithm, the proportion of significantly coexpressed clusters decreases compared with the other two methods. This is expected since coexpression is competing with biological homogeneity.

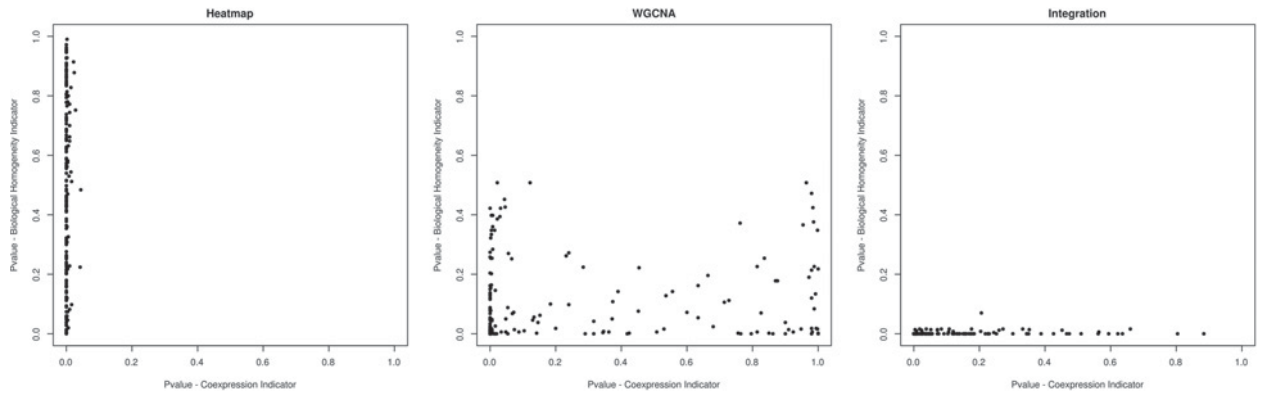


Figure 3 Representation of p-values associated with the coexpression indicator and the biological homogeneity indicator, for the three clustering procedures. Results for the three clustering procedures applied to the chicken expression data are represented: Heatmap (Heatmap), clustering based on coexpression network (WGCNA) and our clustering algorithm (Integration). Whatever the clustering method, to each cluster, is associated a p-value corresponding to the coexpression indicator and a p-value corresponding to the biological homogeneity indicator. P-values are represented in a joint representation, where each point represent a cluster, and p-values associated with the coexpression indicator are represented on the x-axis, whereas p-values associated with the biological homogeneity indicator are represented on the y-axis.

However, the proportion of significantly biologically homogeneous clusters considerably increases (79.50%). This results in a higher proportion of good candidates for interpretation (53.50%).

Note: clusters made up of one single gene are automatically considered as bad candidates. Therefore, as our clustering strategy provided a proportion of these clusters which is not negligible, the percentage of good candidates is mechanically lower.

In conclusion, by integrating biological knowledge into expression data, we manage to obtain a reasonable proportion of clusters, which gather significantly coexpressed and biologically related genes. These clusters are good candidates and their interpretation may lead to reveal new relationships amongst genes.

Clusters interpretation

Clusters obtained by integrating biological knowledge into expression data, and that present interesting properties, are then good candidates for interpretation. In order to associate representative GO annotations with clusters, we choose to apply a classical enrichment testing procedure which consists in fisher's exact tests associated with a correction for multiple testing (Benjamini-Hochberg with a 5% cut-off). The overall impression about the results of the enrichment procedure is the coherence of GO annotations associated with clusters. The enriched GO annotations associated with one cluster are close

Table 2 Results of the case study

	CI	BHI	Both
Heatmap	91.50	13.50	13.50
WGCNA	63.00	68.00	46.00
Integration	53.50	79.50	**53.50**

Results for the chicken data set for the three clustering algorithms: Heatmap classification (Heatmap), clustering based on coexpression network (WGCNA) and our clustering algorithm (Integration). We give the percentage of clusters (%), amongst a given partition, which are significantly coexpressed (CI), biologically homogeneous (BHI) or both coexpressed and biologically homogeneous (Both).

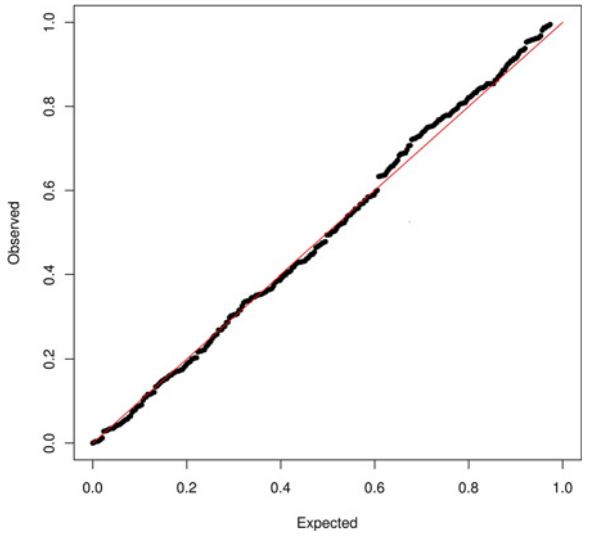


Figure 4 QQ-plot of the p-values associated with the biological homogeneity indicator in the Heatmap clustering. We focus here on the p-values associated with the biological homogeneity indicator of clusters obtained from the Heatmap clustering of chicken expression data. The QQ-plot intersects probabilities that are expected within a uniform distribution (x-axis), with p-values from the biological homogeneity indicator (y-axis).

in the GO hierarchy. This directly conveys the biological homogeneity of gene clusters which is guaranteed by our procedure.

In comparison with the paper by Désert et al. [15], the general and well-known mechanisms implemented in reply to fasting are also highlighted through the enriched annotations of the clusters. In addition, our procedure brings to light new tracks. For instance, a few clusters are associated with Phospholipid and Sphingolipids mechanisms, and whose genes are expressed in fasting chickens, are not described in Désert et al.. These clusters gather several enzymes that are implicated in the hydrolysis of these lipids which results in freeing fatty acids. Then, we think that in chickens, after a certain period of fasting, fatty acids may be consumed from the plasma membrane.

Discussion and conclusion

We propose a new unsupervised gene clustering algorithm which relies on a new distance between genes by integrating biological knowledge into expression data. To do so, we propose a judicious coding that relies on the concept of coexpressed biological function. As a biological function can be assimilated to a set of genes that are involved in the function, we can assimilate a coexpressed biological function to a restriction of the set to coexpressed genes. Naturally, this distance is used to cluster genes.

The properties of gene clusters are then assessed by means of two indicators that we also propose, and which allow to quantify coexpression and biological homogeneity. On the one hand, coexpression is evaluated by an indicator based on correlations between genes. This indicator is purely empirical, but very convenient and easy to interpret. On the other hand, biological homogeneity is measured by an indicator based on Cramér's V coefficient calculated from a matrix which encodes GO annotations. Although this indicator has its limits as biological homogeneity should principally rely on biological interpretation, it happens to be useful to automatically have an idea of the biological interest of gene clusters. In addition, we propose hypothesis testing to enhance these indicators with p-values, in order to verify whether clusters are significantly coexpressed and biologically homogeneous.

To test our clustering algorithm as well as our evaluation procedure, we apply it to both simulated and real data sets. In addition, to position our method we compare it with two gene clustering strategies which are classically used by biologists: Heatmaps and clustering based on coexpression network.

Concretely our methodology shows some limitations as it provides a relatively important proportion of clusters constituted with one single gene. However, it outperforms the other methods: actively integrating biological knowledge into expression data provides partitions with the highest proportion of good candidates. These clusters indeed appears to be good candidates for interpretation as can testify the ones related to Phospholipid and Sphingolipids mechanisms. However an ultimate external biological validation remains to be done, what consists in conducting more advanced biological interpretations.

Appendix

Appendix 1: Range of variation of the coexpression indicator

The coexpression indicator consists in calculating the average of genes correlations within a cluster K_l. Let us recall the calculation of the coexpression indicator(Equation (3)):

$$CI(K_l) = \frac{1}{\frac{card(K_l)(card(K_l)-1)}{2}} \sum_{k|k\in K_l} \times \left(\sum_{k'|k'\in K_l, k'>k} \frac{1}{I} \sum_{i=1}^{I} \left(\frac{G_{ik} - G_{.k}}{S_k} \right) \left(\frac{G_{ik'} - G_{.k'}}{S_{k'}} \right) \right)$$

CI's minimum varies according to $card(K_l)$. In order to obtain a maximum of negative correlations within a K_l, we consider two sub-groups such as intra-group correlation equals 1 and inter-group correlation equals -1. All genes of K_l are equally distributed between both sub-groups.

If $card(K_l)$ is even

In this case, each sub-group is formed by $\frac{card(K_l)}{2}$ genes. The maximum number of negative correlations is equal to $\frac{card(K_l)}{2} \times \frac{card(K_l)}{2}$.

$$CI(K_l) = \frac{[\frac{card(K_l)(card(K_l)-1)}{2} - (\frac{card(K_l)}{2})^2] - (\frac{card(K_l)}{2})^2}{\frac{card(K_l)(card(K_l)-1)}{2}}$$

$$CI(K_l) = -\frac{1}{card(K_l) - 1}$$

If $card(K_l)$ is odd

In this situation, one of the sub-group is constituted by $\frac{card(K_l)-1}{2}$ genes, the other by $\frac{card(K_l)+1}{2}$. The maximum number of negative correlations equals $\frac{card(K_l)-1}{2} \times \frac{card(K_l)+1}{2}$.

$$CI(K_l) = \frac{\left[\frac{card(K_l)(card(K_l)-1)}{2} - \frac{card(K_l)-1}{2} \times \frac{card(K_l)+1}{2}\right] - \frac{card(K_l)-1}{2} \times \frac{card(K_l)+1}{2}}{\frac{card(K_l)(card(K_l)-1)}{2}}$$

$$CI(K_l) = -\frac{1}{card(K_l)}$$

CI is maximum and equals 1 when all genes K_l are perfectly positively correlated.

Appendix 2: Range of variation of the biological homogeneity indicator

Let us recall the calculation of the biological homogeneity indicator(Equation (4)):

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(\sum_{j=1}^{J}\frac{\left(T_{kj} - \frac{T_{k.}T_{.j}}{T_{..}}\right)^2}{\frac{T_{k.}T_{.j}}{T_{..}}}\right)}{T_{..}(card(K_l)-1)}}$$

where T_{kj} equals 1 if the gene k is associated with the biological function j, 0 else wise, $T_{k.}$ is the row margins associated with the gene k.

BHI is minimum and equals 0 when none of the genes of K_l have similar functional signature to such an extend that none of the biological functions is associated with two genes of K_l :

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(\sum_{j=1}^{J}\frac{\left(T_{kj} - \frac{T_{k.}T_{.j}}{T_{..}}\right)^2}{\frac{T_{k.}T_{.j}}{T_{..}}}\right)}{T_{..}(card(K_l)-1)}}$$

$\forall j | T_{kj} = 1,\ T_{.j} = 1$
$\forall j | T_{kj} = 0,\ T_{.j} = 0$

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(T_{k.}\frac{\left(1-\frac{T_{k.}}{T_{..}}\right)^2}{\frac{T_{k.}}{T_{..}}} + (T_{..}-T_{k.})\frac{T_{k.}}{T_{..}}\right)}{T_{..}(card(K_l)-1)}}$$

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(T_{..}\left(1-\frac{T_{k.}}{T_{..}}\right)^2 + T_{k.} - \frac{T_{k.}^2}{T_{..}}\right)}{T_{..}(card(K_l)-1)}}$$

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(\frac{T_{..}^2 - 2T_{..}T_{k.} + T_{k.}^2 + T_{..}T_{k.} - T_{k.}^2}{T_{..}}\right)}{T_{..}(card(K_l)-1)}}$$

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l} T_{..} - \sum_{k\in K_l} T_{k.}}{T_{..}(card(K_l)-1)}}$$

$$BHI(K_l) = 1 - \sqrt{\frac{card(k_l)T_{..} - T_{..}}{T_{..}(card(K_l)-1)}}$$

$$BHI(K_l) = 0$$

BHI is maximum and equal to 1 when all genes of K_l have perfectly similar functional profiles:

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(\sum_{j=1}^{J}\frac{\left(T_{kj} - \frac{T_{k.}T_{.j}}{T_{..}}\right)^2}{\frac{T_{k.}T_{.j}}{T_{..}}}\right)}{T_{..}(card(K_l)-1)}}$$

$\forall j | T_{kj} = 1,\ T_{.j} = card(K_l)\ \&\ T_{.k} = \frac{T_{..}}{cardK_l}$
$\forall j | T_{kj} = 0,\ T_{.j} = 0$
Therefore :

$$BHI(K_l) = 1 - \sqrt{\frac{\sum_{k\in K_l}\left(\sum_{j=1}^{J}\frac{\left(1-\frac{\frac{T_{..}}{card(K_l)}card(K_l)}{T_{..}}\right)^2}{\frac{\frac{T_{..}}{card(K_l)}card(K_l)}{T_{..}}}\right)}{T_{..}(card(K_l)-1)}}$$

$$BHI(K_l) = 1$$

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MV, SL and JP developed the methodology and drafted the manuscript. MV implemented the algorithm. All authors approved the final manuscript.

Acknowledgements
The authors thank Sandrine Lagarrigue, from the Genetic Department of Agrocampus Ouest, for her availability and for letting them use her data. The authors thank the reviewers for their valuable comments.

Received: 4 April 2012 Accepted: 18 January 2013
Published: 7 February 2013

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci* 1998, **95**(25):14863–14868.
2. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4:**Article 17.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25–29.
4. Bauer S, Gagneur J, Robinson PN: **GOing Bayesian: model-based gene set analysis of genome-scale data.** *Nucleic Acids Res* 2010, **38:**3523–3532.
5. Busold CH, Winter S, Hauser N, Bauer A, Dippon J, Hoheisel JD, Fellenberg K: **Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data.** *Bioinformatics* 2005, **21**(10):2424–2429.

6. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data.** *Proteomics* 2007, **7**(13):2162–2171.
7. Yeung MKS, Tegnér J, Collins JJ: **Reverse engineering gene networks using singular value decomposition and robust regression.** *Proc Natl Acad Sci* 2002, **99**(9):6163–6168.
8. Bryan J: **Problems in gene clustering based on gene expression data.** *J Multivariate Anal* 2004, **90:**44–66.
9. Kashima H, Yamanishi Y, Kato T, Sugiyama M, Tsuda K: **Simultaneous inference of biological networks of multiple species from genome-wide data and evolutionary information: a semi-supervised approach.** *Bioinformatics* 2009, **25**(22):2962–2968.
10. Azuaje F, Wang H, Zheng H, Léonard F, Rolland-Turner M, Zhang L, Devaux Y, Wagner D: **Predictive integration of gene functional similarity and co-expression defines treatment response of endothelial progenitor cells.** *BMC Syst Biol* 2011, **5:**46.
11. Li XL, Tan YC, Ng SK: **Systematic gene function prediction from gene expression data by using a fuzzy nearest-cluster method.** *BMC Bioinformatics* 2006, **7**(Suppl 4):S23.
12. Datta S, Datta S: **Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.** *BMC Bioinformatics* 2006, **7:**397.
13. Cramér H: *Mathematical Methods of Statistics (PMS-9)*. New Jersey: Princeton University Press; 1945.
14. Dray S: **On the number of principal components: A test of dimensionality based on measurements of similarity between matrices.** *Comput Stat Data Anal* 2008, **52**(4):2228–2237.
15. Désert C, Duclos M, Blavy P, Lecerf F, Moreews F, Klopp C, Aubry M, Herault F, Le RoyP, Berri C, Douaire M, Diot C, Lagarrigue S: **Transcriptome profiling of the feeding-to-fasting transition in chicken liver.** *BMC Genomics* 2008, **9:**611.
16. Friguet C, Kloareg M, Causeur D: **A factor model approach to multiple testing under dependence.** *J Am Stat Assoc* 2009, **104**(488):1406–1415.

doi:10.1186/1471-2105-14-42
Cite this article as: Verbanck *et al.*: A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. *BMC Bioinformatics* 2013 **14**:42.

CHAPITRE 4

PRISE EN COMPTE DE LA LOCALISATION CHROMOSOMIQUE

Dans ce chapitre, nous présentons un travail collaboratif qui a été effectué avec le laboratoire de génétique animale d'Agrocampus Ouest à l'occasion du doctorat de Marion Ouédraogo. Ce travail nous a conduit à relier la coexpression des gènes à la localisation chromosomique. L'idée motivant le développement d'une telle méthodologie repose sur l'hypothèse suivante : des gènes proches les uns des autres sur le génome peuvent être coexprimés car la structure du génome dans le noyau de la cellule est impliquée dans la régulation de l'expression des gènes. Ainsi, nous avons mis au point un nouvel algorithme qui permet premièrement de définir des régions du génome dans lesquelles les gènes sont à la fois coexprimés et colocalisés et deuxièmement de comparer ces régions. Nous présentons succinctement l'algorithme développé lors de cette collaboration et nous insisterons sur les contributions concrètes que nous avons pu apporter à ce travail.

Ce chapitre inclut l'article :

Ouédraogo, M., Lê, S., Verbanck, M., Diot, C. & Lecerf, F. (2013). Identification of gene coexpression structures by multivariate analyses of expression data and comparison with genome interactions. *Nucleic Acid Research* (soumis)

Sommaire

1 Prise en compte de la localisation chromosomique

Comme nous l'avons exposé section 2.2.2 chapitre 1, des gènes contigus sur un même chromosome peuvent être corégulés, donc coexprimés. C'est pourquoi, il semble naturel d'intégrer une information de localisation chromosomique dans l'analyse de la coexpression des gènes. Nous pouvons ainsi chercher à définir des clusters de gènes colocalisés et coexprimés.

Par ailleurs, en plus de la corégulation de gènes contigus sur le chromosome, nous avons rappelé l'organisation du noyau de la cellule en termes de territoires chromosomiques avec à l'intersection des territoires, les usines de transcription. Dans ces usines de transcription, les chromosomes peuvent interagir. Ainsi, il semble naturel après avoir défini des régions chromosomiques composées de gènes colocalisés et coexprimés de vouloir comparer ces régions afin de mettre en évidence des interactions entre chromosomes.

Dans un second temps, étant donné que la structure des chromosomes dans le noyau affecte l'expression des gènes, l'expression des gènes est en partie une projection de cette structure, nous pouvons imaginer retrouver l'architecture du génome dans le noyau à partir des régions définies et particulièrement de leur interactions.

Nous proposons donc de développer une méthodologie statistique permettant de prendre en compte la localisation des gènes dans l'étude des données d'expression.

1.1 Traduction statistique

Avant de présenter la méthodologie à proprement parler, nous proposons de définir clairement la notion de gènes colocalisés et coexprimés.

1.1.1 Coexpression

Nous nous intéressons de nouveau à la coexpression des gènes (comme dans les chapitres 2 et 3) que nous avions définie comme une forte corrélation positive entre gènes. Cependant, nous avons choisi de définir la coexpression différemment dans cette étude. Nous prenons en compte à la fois la corrélation positive et négative pour définir la coexpression. En effet, comme nous l'avons rappelé (section 2.2.2 chapitre 1), il existe des phénomènes de régulation qui peuvent à la fois activer l'expression de certains gènes et d'inhiber l'expression d'autres gènes d'une même région chromosomique. Par conséquent, deux gènes d'une même région chromosomique, dont l'expression est bien régulée par le même agent, peuvent présenter des profils d'expression corrélés négativement. Ainsi, pour être en cohérence avec la réalité biologique, nous choisissons de définir la coexpression comme une forte corrélation, soit positive soit négative, dans ce cadre.

1.1.2 Colocalisation

En termes statistiques, la colocalisation de gènes est définie par l'intermédiaire d'une fenêtre de k gènes contigus sur un même chromosome. Au sein de cette fenêtre nous considérons que les gènes sont colocalisés. Nous étudions ensuite la coexpression des gènes au sein de chaque fenêtre afin de mettre en évidence des gènes colocalisés et coexprimés. Enfin, pour étudier, sur tout un chromosome, la coexpression de gènes colocalisés, il suffit d'utiliser une fenêtre glissante le long du chromosome.

1.2 Algorithme

Au cours de cette collaboration, nous avons développé un algorithme permettant d'étudier les données d'expression en lien avec la localisation chromosomique. Nous proposons d'en rappeler les grandes lignes car l'écriture de l'algorithme, en prenant en compte les choix méthodologiques, a été une grande partie du travail, bien que son implémentation ait été intégralement réalisée par Marion Ouédraogo sous la forme d'un programme informatique interfaçable avec R au sein d'un package `CoCoMap`. Il s'agit d'une approche modulaire en 3 étapes :

1. étude de la structure de coexpression au sein de gènes colocalisés
   → **définition grossière de régions**
2. étude de la structure de coexpression au sein d'une région chromosomique
   → **affinement de la définition des régions**
3. étude de la structure de coexpression des différentes régions au sein d'un chromosome ou entre les chromosomes
   → **comparaison des régions**

1.2.1 Mise en évidence de gènes colocalisés et coexprimés

La première étape consiste à définir des régions chromosomiques « primitives » dans lesquelles les gènes sont colocalisés et coexprimés. Cela consiste à étudier la structure de coexpression au sein de k gènes colocalisés à l'intérieur d'une fenêtre d'un chromosome. Étudier la coexpression des k gènes colocalisés revient à étudier la structure de corrélation des k profils d'expression. Pour cela, nous proposons de réaliser l'ACP du tableau croisant les sujets et les k gènes de la fenêtre. Cette ACP est associée à un test d'hypothèse sur la première valeur propre, qui fournit une probabilité critique permettant de quantifier la significativité de la structure de coexpression des k gènes colocalisés. Enfin, la fenêtre de k gènes est glissée le long de chaque chromosome.

Remarque : dans le chapitre 3, nous nous sommes intéressés à la significativité de la coexpression globale de clusters de gènes. Pour cela nous avons proposé un indicateur de coexpression basé sur les corrélations des gènes du cluster. Dans ce cadre, nous proposons une mesure différente, basée sur la première valeur propre. Cela est directement lié à la définition de la coexpression. En effet, pour des raisons d'adéquation avec la réalité biologique que nous cherchons à prendre en compte, la coexpression peut être définie comme une forte corrélation positive (chapitre 3) ou comme une forte corrélation (chapitre 4).

In fine, la première étape de la procédure fournit une probabilité critique pour chaque fenêtre le long du chromosome. Les valeurs propres et/ou les probabilités critiques obtenues sont représentées le long du chromosome sous forme d'un graphique que nous appelons *autovariogramme* (figure1).

Remarque : l'algorithme a donc un paramètre de réglage qui est la taille de la fenêtre (k).

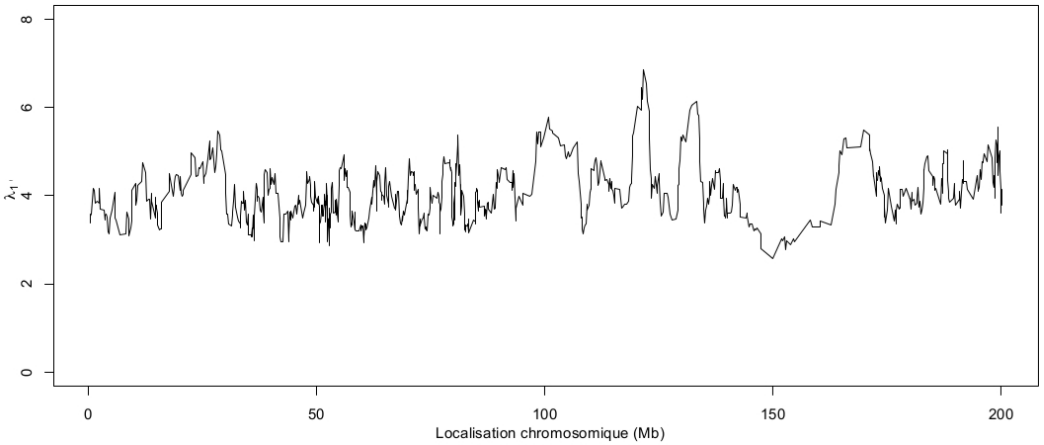


FIGURE 1 – Exemple d'autovariogramme pour un chromosome avec en abscisse la position du premier gène de la fenêtre sur le chromosome (en mégabase (Mb) : unité représentant un million de bases azotées d'ADN) et en ordonnée la première valeur propre de l'ACP sur les gènes de la fenêtre.

1.2.2 Définition de régions chromosomiques

A partir de l'autovariogramme, nous pouvons définir des régions chromosomiques dans lesquelles les gènes sont significativement coexprimés. Une région est donc définie comme un ensemble de fenêtres contiguës dont les gènes sont coexprimés. Pour identifier une région, le chromosome est scanné à travers son autovariogramme et la première fenêtre significativement coexprimée (probabilité critique associée à la première valeur propre de la fenêtre en question, inférieure à un seuil) correspond au début d'une région. La dernière fenêtre d'une région correspond ensuite soit à la dernière fenêtre du chromosome si toutes les fenêtres sont significatives entre le début de la région et la fin du chromosome, soit à la dernière fenêtre le long du chromosome qui présente une probabilité critique inférieure au seuil. Cette étape permet de définir des régions préalables qui sont ensuite affinées.

Une fois que les régions sont grossièrement définies, elles sont redéfinies. Les fenêtres qui font partie d'une même région sont testées afin de déterminer si elles présentent bien une structure de coexpression commune. Pour cela, nous utilisons le coefficient Lg (Escofier et Pagès, 2008) qui mesure la similarité entre deux groupes de variables. Le Lg est ainsi calculé entre le groupe de variables formé par tous les gènes de la région et chaque groupe constitué par les variables d'une fenêtre. Ce coefficient est ensuite associé à un test d'hypothèse qui permet de ne conserver que les fenêtres significativement structurée comme la région globale. De plus, les fenêtres situées aux abords de la région sont également testées pour s'assurer de leur non appartenance à la région et ainsi redéfinir si nécessaire les limites de la région.

Nous modérons ensuite cette notion de région. Effectivement, une région est définie comme un ensemble de fenêtres contiguës dont les gènes présentent des structures d'ex-

pression similaires. Ainsi, une région revient à considérer un ensemble de gènes qui sont consécutifs sur le chromosome. Cependant, tous les gènes définis comme appartenant à une région ne sont pas nécessairement coexprimés, à tel point que les gènes d'une région peuvent être classés en 2 sous-ensembles de gènes coexprimés. Cela revient donc à considérer deux sous-régions qui sont superposées. Ainsi, au sein d'une région qui cache deux sous-régions superposées, non seulement la première composante principale représente une structure de coexpression, mais également la seconde, chacune des deux composantes principales représentant la structure de coexpression d'une sous-région.

En conclusion, une fois toutes ces étapes de redéfinition des régions effectuées, nous disposons de régions. Chaque région est représentée par les S premières composantes principales de l'ACP du tableau croisant les sujets et les gènes de la région, S étant spécifique à chaque région et correspond au nombre de composantes significatives de la région. Nous souhaitons maintenant comparer ces régions entre elles.

1.2.3 Comparaison des régions

Les deux premières étapes consistent à identifier des régions chromosomiques qui sont constituées de gènes de fait colocalisés et coexprimés. Nous proposons une troisième étape dans notre méthodologie qui consiste à comparer les régions entre elles afin d'identifier les régions présentant une structure de coexpression similaire. Pour cela, chaque région est représentée par ses S premières composantes principales elles-mêmes représentant des sous-régions superposées de gènes colocalisés et coexprimés. Pour comparer les régions entre elles, sur la base de leur structure de coexpression, nous proposons de réaliser une classification ascendante hiérarchique afin de classer les régions. Une matrice de distance est obtenue sur la base des S premières composantes principales de chaque région.

In fine, l'algorithme fournit des clusters de gènes colocalisés et coexprimés sous forme de régions chromosomiques ainsi que des groupes de régions chromosomiques exposant la même structure de coexpression. Les régions ainsi définies qui sont des clusters de gènes peuvent être interprétés comme classiquement au moyen de tests d'enrichissement par exemple. De plus, nous pouvons tenter de retrouver l'architecture du génome dans le noyau à partir des régions définies et de leurs interactions.

2 Apports concrets dans la mise au point de la méthodologie

Ce travail collaboratif était pleinement centré autour du travail de doctorat de Marion Ouédraogo, biologiste de formation, qui a implémenté la totalité des méthodologies statistiques développées. Cette collaboration s'est déroulée sur deux années d'allers-retours représentant réellement plusieurs semaines de travail commun. *In fine*, cette collaboration fut une expérience très enrichissante de **bio**statistique en étant parfaitement à l'interface

entre les deux disciplines et entre les deux types d'interlocuteurs généticiens et statisticiens.

Nous insistons donc dans cette section le rôle concret que j'ai joué dans cette collaboration, mes contributions se situent à plusieurs niveaux.

2.1 Intermédiaire entre génomique et statistique

Un des défis de cette collaboration a été de traduire les problématiques biologiques émises par les généticiens en problématiques statistiques. En effet, il est essentiel d'avoir une compréhension globale des choses à la fois en biologie et en statistique pour être en mesure de développer puis de valider les méthodologies et pour s'assurer de la concordance entre les méthodologies et les principes biologiques qui les ont motivées. J'ai donc joué pleinement ce rôle d'intermédiaire afin de permettre aux statisticiens de comprendre les problématiques soulevées par les généticiens.

Une fois les problématiques clairement énoncées, il a fallu construire une stratégie statistique afin d'y répondre au mieux. J'ai donc été impliquée dans la prise de décision. Comme nous l'avons déjà évoqué, l'implémentation de l'algorithme a été intégralement réalisée par Marion Ouédraogo. Cependant, une fois les choix méthodologiques définis, autrement dit, une fois la traduction des problématiques biologiques en traitements statistiques effectuée, il a été nécessaire d'écrire véritablement l'algorithme synthétisant tous ces choix, ce que j'ai réalisé pour permettre l'implémentation. Plus précisément, il m'a fallu véritablement décortiquer l'algorithme que nous avons développé afin de permettre à chaque étape aux généticiens de comprendre les méthodologies mises au point et à Marion Ouédraogo d'implémenter l'algorithme.

Enfin, un des aspects importants de cette collaboration a été d'initier les biologistes à la validation de méthodologies statistiques par l'intermédiaire de simulations. J'ai donc proposé des simulations afin de tester et valider la méthode. Ce point a été particulièrement important et c'est celui que nous développerons un peu plus en détail par la suite.

2.2 Validation de la méthode : mise au point d'un plan de simulations

2.2.1 Principe des simulations

Pour comprendre la structure des jeux de données que nous simulons, prenons un exemple concret dans lequel nous considérons 2 chromosomes constitués de 450 gènes au total (figure 2). Dans le premier chromosome, nous pouvons identifier 2 régions (nommées G1 et G2) de gènes colocalisés et coexprimés, tandis que nous en avons 4 (nommées G3, G4, G5 et G6) dans le second chromosome. Chaque couleur dans les régions symbolise une certaine structure de coexpression. Ainsi, les régions G3 et G4, situées sur le chromosome 2, présentent une même structure de coexpression et les régions G2 et G6 présentent la même structure de coexpression, bien qu'elles soient situées sur des chromosomes différents. Enfin, la région G5 du chromosome 2 est

partiellement superposée aux régions G4 et G6. Ainsi à travers ces simulations, nous couvrons l'ensemble des situations théoriques.

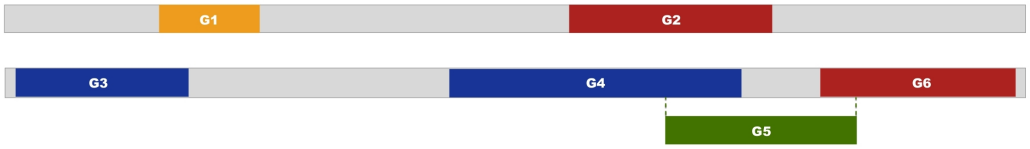


FIGURE 2 – Schéma d'un plan de simulation.

Une région simulée est définie par plusieurs paramètres :

- **la taille de la région** qui le nombre total de gènes entre le premier gène et le dernier gène significativement coexprimés avec les autres gènes de la région
- **le nombre de gènes coexprimés** parmi l'ensemble des gènes de la région
- **la densité de gènes coexprimés** qui est le rapport du nombre de gènes coexprimés sur la taille de la région

Une fois la structure globale définie, nous construisons une matrice de corrélation théorique pour tous les gènes qui reflète la structure globale. Des jeux de données sont ensuite simulés au moyen du package `mvtnorm` (Genz *et al.*, 2013), qui permet de simuler des données aléatoires pour un groupe de variables. Les données sont générées à partir d'une loi normale multivariée de telle sorte que la structure corrélation globale des données calque la structure de corrélation théorique définie.

Pour interpréter les résultats des simulations, nous nous sommes basés sur plusieurs indicateurs classiques :

- précision : rapport entre le nombre de gènes simulés coexprimés et le nombre de gènes identifiés comme coexprimés
- rappel : rapport entre le nombre de gènes simulés coexprimés et identifiés comme coexprimés et le nombre de gènes simulés coexprimés
- mesure F : combinaison du rappel et de la précision qui tend vers 1 lorsque la procédure est efficace

De plus, nous proposons une représentation graphique des résultats de simulations qui donne une bonne vision des performances de la méthode. Cette représentation est proposée figure 3 pour un exemple de simulation suivant le plan présenté figure 2. Au

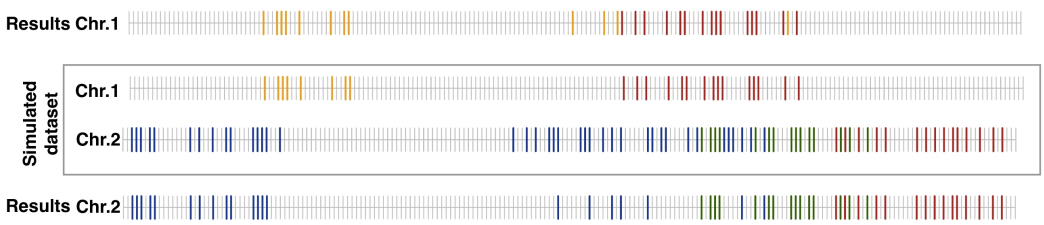


FIGURE 3 – Comparaison des régions simulées et des régions identifiées par l'algorithme `CoCoMap`.

centre, nous retrouvons les données simulées composées de six régions réparties sur les deux chromosomes. Chaque bâton représente un gène : s'il est gris cela signifie que le gène n'est coexprimé avec aucun autre gène, s'il est coloré il appartient à la région correspondant à la couleur. En haut (pour le chromosome 1) et en bas (pour le chromosome 2), nous trouvons les résultats fournis par l'algorithme avec la même nomenclature. Grâce à cette représentation, nous pouvons constater que l'algorithme permet globalement de retrouver les régions et les interactions entre régions.

Les simulations nous ont permis de valider la méthode sur plusieurs points, nous présentons deux aspects de validation.

2.2.2 Influence de la taille de la fenêtre de colocalisation (k)

D'une part, les simulations ont permis de vérifier l'influence de la taille de la fenêtre. En effet, un des paramètres de réglage de la méthode est le choix de la taille de la fenêtre de k gènes colocalisés. Pour analyser l'influence de la taille de la fenêtre de colocalisation, nous simulons des régions chromosomiques en faisant varier la taille de la région (entre 0 et 50) avec une densité constante et égale à 0.5. En conséquent, le nombre de de gènes coexprimés est égal à la moitié de la taille et varie entre 0 et 25. Les différents jeux de données simulés ont été analysé avec `CoCoMap` en considérant des tailles de fenêtre de 5, 10, 15, 20, 25, 30 et 35 gènes. D'après la figure 4, à partir d'une taille de 10 gènes, les performances sont comparables, ce peu importe la taille de la région. Néanmoins, le meilleur compromis correspond aux tailles de fenêtre de 15 et 20 gènes.

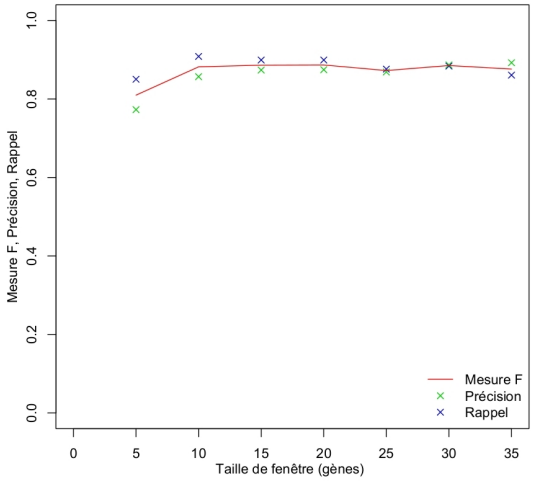


FIGURE 4 – Précision, Rappel et mesure F en fonction de la taille de la fenêtre, toutes tailles de région (entre 0 et 50) confondues.

2.2.3 Influence de facteurs intrinsèques aux données sur l'identification des gènes coexprimés et colocalisés

D'autre part, nous souhaitons tester l'influence du niveau de corrélation des gènes coexprimés de la région, de la densité et du nombre de sujets. Pour cela, nous avons simulés des régions de 160 gènes comprenant 15 gènes coexprimés. Nous faisons ensuite varier :

- le niveau de corrélation absolue des gènes coexprimés de la région (entre 0 à 1)
- la densité (entre 0 et 1)
- le nombre de sujets (entre 10 et 100)

Nous analysons ensuite les performances de la méthodologie (figure 5). A partir d'un niveau de corrélation absolue de 0.4, d'une densité de 0.5, de 20 sujets, ce qui est réaliste, les performances de la méthodologie sont très satisfaisantes et comparables avec les performances obtenues pour des valeurs supérieures de ces trois paramètres.

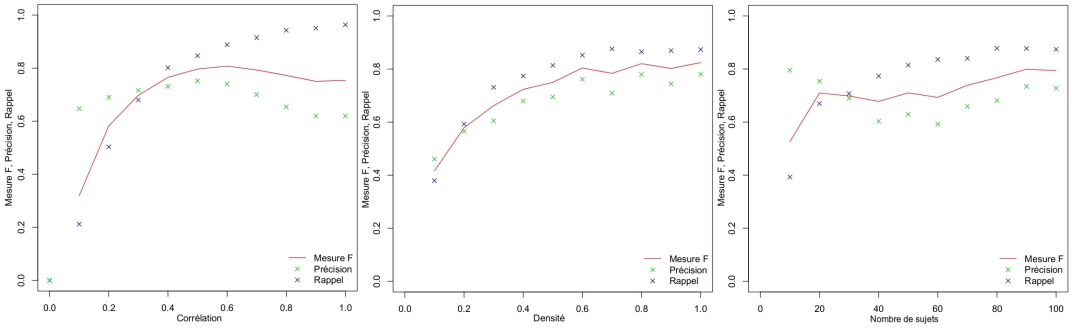


FIGURE 5 – Précision, Rappel et mesure F en fonction du niveau de corrélation absolue (gauche), de la densité de gènes coexprimés (centre) et du nombre de sujets.

Ces résultats sont très encourageants et illustrent le comportement et les performances très satisfaisantes de l'algorithme.

L'algorithme d'étude des données transcriptomiques en lien avec la localisation chromosomique est présenté dans l'article suivant :
Ouédraogo, M., Lê, S., Verbanck, M., Diot, C. & Lecerf, F. (2013). Identification of gene coexpression structures by multivariate analyses of expression data and comparison with genome interactions. *Nucleic Acid Research* (soumis)

Identification of gene coexpression structures by multivariate analyses of expression data and comparison with genome interactions (Ouédraogo *et al.*, 2013)

Résumé : L'architecture du génome et les interactions entre chromosomes jouent un rôle majeur dans la régulation de l'expression des gènes. Des méthodes précises de biologie moléculaires permettent de cartographier ces interactions entre chromosomes (Hi-C, 3C...). Nous proposons, dans cet article, une méthode computationnelle originale, pour identifier les gènes colocalisés et coexprimés, qui est basée sur l'hypothèse suivante : étant donné que la structure du génome dans le noyau affecte l'expression des gènes, l'expression des gènes est en partie une projection de cette structure. Pour cela, nous avons utilisé des méthodes d'analyse multidimensionnelle exploratoire pour mettre en évidence de tels groupes de gènes. La méthode a permis d'identifier la plupart des clusters de gènes coexprimés et colocalisés dans des données simulées, même avec un très petit nombre de sujets et avec un niveau de corrélation faible. Les structures de coexpression identifiées par notre méthode appliquée à des données transcriptomiques coïncident plutôt bien avec les zones d'interaction identifiées par Hi-C. Notre méthode a détecté de nombreuses régions de coexpression, qui peuvent être synthétisées dans 4 structures de coexpression. Des interactions physiques (Hi-C) ont été identifiées pour 63 % des gènes coexprimés dans les 4 structures de coexpression, notons que cette proportion est particulièrement élevée pour l'une des structures. Les chromosomes présentant un grand nombre de régions de coexpression sont localisés près du centre du noyau et les chromosomes présentant un petit nombre de régions sont localisés à la périphérie du noyau d'après les études FISH.

Identification of gene co-expression structures by multivariate analyses of expression data and comparison with genome interactions

Marion Ouédraogo[1,2,*], Sébastien Lê[3,4], Marie Verbanck[3,4], Christian Diot[1,2], Frédéric Lecerf[1,2].

[1] UMR1348, INRA, Rennes, F-35000, France
[2] UMR1348, Agrocampus OUEST, Rennes, F-35000, France
[3] UMR6625, Agrocampus OUEST, Rennes, F-35000, France
[4] UMR6625, CNRS, Rennes, F-35000, France

* To whom correspondence should be addressed. Tel: +33 (0)2 99 23 48 54 64; Fax +33 (0)2 23 48 54 70; Email: marion.ouedraogo@rennes.inra.fr

ABSTRACT

The genome architecture and interactions between chromosomes play a major role in the regulation of gene expressions. Some acute methods of molecular biology allow the mapping of the interactions between chromosomes (Hi-C, 3C…).
We report here an original computational method to map co-located and co-expressed genes based on the following hypothesis: as the structure of the genome within the nucleus affects gene expressions, gene expressions are in some part a projection of this structure. We applied multivariate exploratory approaches of statistics to evidence such group of genes among the genome.
The method identified most of the simulated clusters of co-expressed and co-located genes, even with few samples or very low correlation levels. The comparison of the interactions identified by Hi-C to the co-expression structures identified by our method on transcriptomics data revealed groups of co-expressed genes around interacting regions. The method detected numerous regions of co-expression for 4 structures of co-expression. Physical interactions were found for 63 % of the co-expressed genes, particularly for one structure of co-expression. Chromosomes with numerous regions of co-expression are located to the center of the nucleus, and chromosomes with few regions of co-expression are located near the nuclear periphery according to FISH studies.

INTRODUCTION

In the nucleus, genomic DNA is tightly packaged and organized into higher-level structures required for proper genome function (1, 2)This chromatin conformation is highly dynamic, modified by several biological processes and was recently found to play an important role in the regulation of gene expression(3–5). Several studies have shown that such processes involve physical contact and interaction between distantly located genomic elements within and between chromosomes (6–8). Such examples concern the interaction of the Igf2/H19 locus with its enhancer through long-range

loop (9–12). It is now clear that physical and spatial association between distant genomic elements is an important mechanism of gene regulation (1, 13–15) .

Several techniques were recently developed to examine chromatin structure at high-resolution (for review, see 16), including Chromosome Conformation Capture (3C) (17), Circular Chromosome Conformation Capture (4C) (18, 19), Chromosome Conformation Capture Carbon Copy (5C)(20), Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) (21), the technology developed by Duan et al. (22), and Hi-C (23). These techniques combine various high-throughput approaches and produce large datasets. These data broadly define the three-dimensional conformation of chromatin. However, these methods are known to be labour intensive and can hardly be conducted in large-scale studies including several samples, i.e. individuals, tissues, physio-pathological status…

Here we propose another approach based on the following hypothesis: as the structure of the genome within the nucleus affects gene expressions, gene expressions are in some part a projection of this structure. We therefore developed an original statistical method based on multivariate analyses to investigate the structure of co-expression between genes taking the spatial dimension into account. This method is computational and takes advantage of gene expression data that are already available and correspond to a large number of situations (24, 25)

MATERIAL AND METHODS

Datasets simulations

Different data sets were simulated using the mvtnorm R-package that simulates data according to a given structure of correlation.

In a first series of data, chromosomal regions including 50 genes were simulated. For each region, 3 to 25 genes were defined as co-expressed. The regions were varying according to the co-expressed genes locations and/or organisations. The simulations were repeated 50 times, resulting in a series of 600 datasets.

A second series of data was simulated with i) density of co-expressed genes within the 50 genes region varying from 0.4 to 0.9, ii) correlation level between co-expressed genes varying from 0.3 to 0.9 and iii) number of samples varying from 20 to 100. Each combination of these conditions was simulated 10 times, resulting in a series of 2100 datasets.

A third series of data was simulated including expression data for 410 genes located in two chromosomes and for 50 different samples. Six regions of co-expressed and co-located genes were simulated including two co-expressed regions within the same chromosome, two co-expressed regions between the two chromosomes and one independent region that overlaps two other regions and finally, one region whose expression structure is totally independent from the others. The correlation between the co-expressed genes was equal to 0.5.

Gene expression data

A dataset corresponding to transcriptomic analyses of 12 normal human tissues (26) using microarrays was retrieved from Gene Expression Omnibus (GSE803).
Before analyses, these expression data were pre-processed. The chromosomal locations of the genes corresponding to the probes spotted on the microarrays were retrieved using the Biomart tool(27). The transcriptomic datasets were then filtered to exclude all data from the probes with neither gene ID, nor genomic location, or with “random” or “unknown” chromosomes locations. Data from different probes corresponding to the same gene were reduced to one synthetic data corresponding to the first principal component of data from the subset of probes.

Multivariate exploratory approach

We have developed a 3-step method to determine the regions of co-expressed and co-located genes using the pre-processed data.
The first step consisted in studying the co-expression structure among co-located genes which were defined as neighbouring genes of a same chromosome (Fig1., step 1). To this aim, a window of n co-located genes was slid gene by gene along each chromosome. The expression structure of the subset of genes in the window was studied using Principal Component Analysis (PCA). For each of the windows co-location, a normalized (PCA) was performed. In association with PCA, we used hypothesis testing to determine to what extend the first principal component represented a co-expression structure. The principle of hypothesis testing lies in comparing the observed co-expression structure, which is quantified by the first eigenvalue λ1, with random co-expression structures. Thus, random gene expressions were obtained by permuting the original n gene expression data. PCA was computed on the random gene expression data that provided an expected first eigenvalue, λ1exp. The permutation trials were carried out 500 times to compute a p-value. Ultimately, the first step of the procedure provided an observed eigenvalue, λ1obs, and p-values for all windows along each chromosome.
The second step consisted in discerning regions of co-expressed genes (Fig. 1, step 2a): a region was defined as a set of contiguous and overlapping windows whose genes exhibited similar expression structures. To reduce computation time, chromosomal regions including windows sharing the same structure were first preselected using λ2/λ1 ratio (Fig. 1, step 2a). Contiguous windows with a low λ2/λ1 value, i.e. with a high λ1 value, share the same structure of co-expression. Conversely, when λ2/λ1 ratio tends towards 1, there are two independent structures within the observed window. The chromosome was scanned and the first encountered significant window defined the start of a region (step 1). The end of the region was defined according to three different parameters: the last screened window had a λ2/λ1 ratio above a defined threshold (empirically set to 0.75) and/or a p-value lower than the associated threshold and/or was really the last window at the end of the chromosome. Once regions were “roughly” defined as previously described, they were refined (Fig. 1, step 2b). Windows that were part of the same region were tested for common co-expression structure using Multiple Factor Analysis (MFA) in association with hypothesis testing of the measurement coefficient denoted Lg. Lg quantifies the adequacy of the co-expression structure of a window to the co-expression structure of the region. In association with MFA, we used hypothesis testing to

determine to what extend the co-expression structure of a window was similar to the co-expression structure of the region. Random structure of expression for a window was thus obtained by permuting the original n gene expression data. MFA was computed after replacing the original window by the permutated window. The random structure provided an expected Lg coefficient. The permutations trials were carried out 500 times to compute a p-value.

As depicted in figure 1, some genes belonging to a co-expression structure can be located outside the pre-selected region (the λ2/λ1 ratio threshold might be too stringent in some cases). To overcome this apparent discrepancy, the flanking windows were included in the analysis in a second step. This process was iteratively extended until the Lg p-values became insignificant.

To identity the co-expressed genes, identification of the expression data that contribute to the structure of expression is needed. Statistically, this corresponds to identify the significant variables that describe a PC. For each region, the number of significant PC was defined using the general cross-validation criterion (Husson et al. 2010a). The number of significant PC was limited to a maximum value of 5. The significant variables that describe the PC were defined using the method of Husson et al. (2010b). The probability of a variable to be associated to one PC was corrected for multiple testing using the Benjamini-Hochberg method.

At the third step the similarities of co-expression structures between regions within or between chromosomes were assessed at the genome scale. To this aim, we used Hierarchical Ascendant Clustering (HAC) to classify regions through their significant principal components. Using the Ward's method, HAC was performed on a distance matrix between all principal components computed as $1-cor(PC_i,PC_j)^2$. The tree was then automatically cut to select a limited number of clusters so that the selected regions present the more similarities within clusters but the more differences between clusters. The principal structures of co-expression corresponded to the structures of co-expression of the clusters.

At the end of these processes, datasets were constructed including the co-expressed genes, their genomic location, regions of co-expression and associated PC, and cluster of co-expressed regions.

Hi-C interaction and co-located genes co-expression data comparison

The Hi-C method allows identification of chromatin interactions across an entire genome. In our study, physical interactions for every gene included in the transcriptomics dataset were selected from the Hi-C dataset. The genes that were not represented in the Hi-C dataset were considered as genes without interaction. Interaction data corresponding to these genes (i.e. genes without and with Hi-C interactions) were compared to co-expression data. All the genes that interacted according to the Hi-C data and that were co-expressed were considered as genes that share common relations. The genes that were only found co-expressed or only interacting were considered as genes with specific relations according to the method. We used hypothesis testing to determine to what extend the two methods identified relations between the same genes. For that purpose, an expected number of common and specific relations was obtained by permutations of co-expression data.

RESULTS

To identify regions of co-expressed and co-located genes, we have integrated multivariate exploratory approaches of statistics. These allowed i) the description of local structures of gene expression using PCA, ii) the definition of larger regions with the same expression structure using MFA, and iii) the definition of interactions between these regions, i.e. higher order structures of expression, using a classification approach (see Figure 1 and methods section for details).

Effects of the length of the window and the number of co-expressed genes on the description of local structures of co-expression

Description of local structures of gene expression within a window was first realised. To test the effect of the length of this window, a first set of simulated data (see methods) was analysed with 7 different windows varying from 5 to 35 genes (Fig. 2A). Whatever the length of the window, the method identified the regions of co-expressed and co-located genes in more than 82.5% of the simulated datasets. This efficiency was slightly better with smaller windows. On average, 87% (median) of the simulated co-expressed genes were identified. These were defined as true positives genes (TP). Conversely, 30% (median) of the genes identified by the method were not simulated as co-expressed. These were defined as false positive genes (FP). The length of the window had only a slight effect on the percentage of FP identified.

The effect of the number of simulated co-expressed genes within the regions was next analysed. A strong effect of the number of co-expressed genes was observed on the efficiency of the method, especially with small numbers of co-expressed genes (Fig. 2B). Indeed, the percentages of TP detected were the lowest, 0% to 33% (according to the length of the window) and 59% to 63%, when respectively 3 or 5 co-expressed genes were simulated. The lowest TP percentages were found with the largest windows. With 7 or more simulated co-expressed genes the TP percentages reached a plateau (from 74% to 100%), without effect of the length of the window. A symmetric effect was observed on the percentages of FP identified. They were the highest (from 51% to 99%) when 3 or 5 co-expressed genes were simulated. Increasing the number of co-expressed genes from 3 to 7 or more decreased the FP percentages from 100% to 31%. A slight effect of the length of the window was also observed on the FP percentages, especially with the smallest (5 genes) window. Few differences were observed with larger windows although the lowest FP percentages were observed with the 15 and 20 genes windows. Thus, further analyses were conducted with a 15 genes window.

Effects of gene density, correlation levels and number of samples

The effects of density of co-expressed genes within a window, correlation level between gene expressions and number of samples were also tested using a second set of simulated data. These data were analysed with a sliding window of 15 genes and a significant threshold of 5 %.

As shown in figure 3A, more the number of samples was important, more efficient the method to identify TP. However, the number of samples required was reduced when correlation of expression between samples increases: at least 60 samples were required with a 0.4 correlation value to detect 100% TP and only 20 samples with a 0.9 correlation value.

Similar patterns were not observed on FP percentages (Fig. 3B). For the lowest correlation values (0.4 to 0.7), a slight decrease of FP percentages was observed (from 20-25% to 15-20%) when the number of samples increased. However, with 0.8 and 0.9 correlation values the FP percentages increased from 20% to 30% with 40 samples and to 40% with 50 samples, respectively, and then decreased with higher number of samples to values similar to those observed with lower correlation values.

The density of co-expressed genes within the 50 genes region had only a marginal effect on TP percentages (Fig. 3C), most of the sample numbers leading to nearly 100% of TP genes detected, when 20 and 40 were excepted (90% and 97%, respectively).

As observed in figure 3B, FP percentages increased from 20 to 40 samples (20% to 30%, respectively) and then decreased regularly when the number of samples increased (leading to 10-18% for 100 samples, Fig. 3D). However, few differences were observed according to the density of co-expressed genes.

Density of co-expressed genes together with correlation level (Fig. 3E) had nearly no effect on the TP percentages, near 100% for all correlation values, except for 0.3 to 0.4 values with slightly lower percentages (92-96%).

Similarly and whatever the correlation value, density had nearly no effect on the FP percentages (Fig. 3F). However, these percentages increased from 20 to 30% as correlation values increased from 0.3 to 0.9.

Altogether, these simulations showed that these 3 parameters have different effects on the TP and FP percentages of co-expressed genes detected. The density of co-expressed genes does not have a major effect, whereas the number of samples and especially the correlation level drastically impact the TP and even more the FP percentages. Datasets with a 0.5 correlation value presented the lowest FP percentage but a high TP percentage even with few samples. Thus, the datasets for the following analyses were simulated with a 0.5 correlation value.

Detection of co-expression structures in simulated data

The efficiency of the method to detect whole-genome co-expressed regions, i.e. higher order structures of expression, was tested using a third series of data simulated according to the previous results. These datasets included expression data for 410 genes in 50 samples, with a global 0.5 correlation value between the co-expressed genes. Groups of co-expressed genes were distributed

between two chromosomes. Six co-expression regions were simulated (Fig. 4) including one independent region (yellow), two co-expressed regions within the same chromosome (blue), two co-expressed regions between the two chromosomes (red) and one independent region that overlap two other regions (green). The dataset was analysed with a 15 genes sliding window and a significant threshold of 5 %. As expected, the method clustered co-expressed genes and identified the two co-expressed regions (blue and red) in 98% of the dataset. The independent region (yellow) was correctly identified and clustered in 99 % of the dataset. The method also identified the overlapping region (green) in 79 % of the datasets.

Detection of co-expressed regions in experimental expression data and comparison with Hi-C interaction data

Finally, our method was applied to experimental expression data and the results were compared to those obtained from physical interaction analyses.

Expression data corresponding to 12 different human tissues (GSE803) were retrieved from Gene Expression Omnibus. They involved 20061 different genes. For 18053 of these genes, interaction data were available in Hi-C data published by Lieberman-Aiden et al.(23) For the 2008 other genes, no interactions were found in Hi-C data. They were considered as genes with no physical interactions. Co-expression structures within the 20061 genes were analysed with our method. Interactions and co-expressions for these 20061 genes were then compared within each chromosome or between two different chromosomes (supplementary table 1). Common relation between two genes was defined when they were found co-expressed and physically interacting.

Altogether, 63% of the genes (12620/20061) presented common relations in the two datasets. However, 30% of the genes that interact in Hi-C data (5433/18053) were not found co-expressed, and 24% of the genes that were found co-expressed (3925/16545) do not interact with the same genes in Hi-C data (supplementary table 1). The number of common relations between the two dataset was significant. Intra- and inter-chromosomal relations of co-expression and physical interactions between the 759 co-expressed regions were drawn together (Fig. 5). These relations of co-expression corresponded to four structures of co-expression. Most of the regions (664, 88%) were involved in relations of both co-expression and physical interactions. The common relations concerned 7599 (44%) of the physical interactions found between two regions, co-expressed or not. However, these interactions were not distributed randomly in the genome (Fig. 5B-5D and supplementary Table 1). Indeed, in the first structure 24% of the relations of co-expression were also found with physical interactions representing 72% of the physical interactions found between two co-expressed regions (Fig. 5B, 5C). Conversely, in the second, the third and the fourth structures only 9%, 2% and 2% of the relations, respectively, were found with physical interactions (Fig. 5D, 5E). Common relations involved mainly chromosomes 1, 17, 19, 22 and X, corresponding to 36% of the interactions found between co-expressed regions. Chromosomes 4, 18 and Y were involved in fewer common relations

corresponding to less than 1% of the interactions found between co-expressed regions. Interestingly, regions of co-expression that belonged to the first structure presented more common relations (5400 common relations) than other structures (respectively, 1566, 234 and 299 common relations). However, the four structures corresponded respectively to 30%, 24%, 26% and 20% of the physical interactions found between co-expressed regions. When analysing the number of co-expressed regions compared to the number of regions with physical interactions for each chromosome, some differences were also observed (supplementary Fig. 1 and Table 2). For examples, chromosomes 19, 22 and X presented more regions of co-expressed genes and more regions with physical interactions, and conversely, chromosomes 4, 18 and Y presented fewer regions of co-expressed genes and fewer regions with physical interactions.

DISCUSSION

Recent developments of molecular biology methods have highlighted the role of genome architecture and interactions between chromosomes in the regulation of gene expression. According to these results, we hypothesized that co-expressed and co-located genes should in some part involve genes that physically interact. We therefore integrated multivariate exploratory approaches to investigate the structure of co-expression between genes using transcriptomics data and taking into account the spatial dimension, meaning gene locations, and compared them to Hi-C interaction data.

The first step of the method allowed the description of local structures of co-expression within a window slide along the genome. By definition, PC summarises the information contained by several correlated variables. Thus, we used PCA to identify the co-expressed genes that corresponded to the variables describing the PC.

We first addressed the efficiency of the method according to the length of the window and the number of co-expressed genes. As observed in figure 2, the length of the window did not significantly affect the number of true and false positive genes and true positive regions identified. Conversely, a strong effect of the number of co-expressed genes was observed on the efficiency of the method (i.e. increasing the percentage of FP), especially with small numbers of co-expressed genes (Fig. 2B). In fact, this small number effect is not surprising because the definition of a significant PC requires a sufficient number of co-expressed genes inside a window, whatever the length of the window. Nevertheless, these tests allowed us fixing the length of the window (i.e. 15 genes) to further analyse the effects of different parameters. Another problem with the determination of the length of the window lies in the fact that groups of co-expressed genes can be larger than the window. Using the MFA method to compare the co-expression structures of contiguous and overlapping windows overcame this problem.

Gene density, correlation levels and the number of samples are 3 parameters of almost importance using expression data. Indeed, gene density varies according to different regions and chromosomes of a genome, expression data are sometimes recorded from very few samples and the correlation level between genes also varies from 0 to 1 depending on the biological contexts studied. Our results on simulated data have shown that gene density did not significantly affect the detection of true

positive genes and had a slight effect on false positives. However, the number of samples and the correlation levels affected both the percentages of TP and FP detected but in different ways. For true positives, the combination of both parameters had to fulfil a threshold: a low correlation level needed a large number of samples and conversely, a high correlation level required a small number of samples. For the false positive, the scheme was quite different since the PCA method associates numerous genes with a low correlation value to a structure of co-expression when using data with a strong global correlation value. Indeed, some variables increase occasionally the level of summarized information for a PC, leading the PC to the first position. Those variables are correlated to each other's but not closely correlated to the PC. Taking this property into account, we expected a relatively high number of false positive genes for the simulated datasets. The significance threshold has to be adjusted due to spurious correlation of the variables to the PC, depending on the initial level of correlation in the dataset. We can also observe that the false positive values were the highest with a medium number of samples (between 40 and 60) and a very high level of correlation. Altogether, it appears that the number of samples and the intrinsic correlation level have the largest impact on the efficiency of the method to identify co-expressed and co-located genes.

The last validation step was to test the efficiency of the method to identify co-expressed structures at a genome scale. When applied on data simulated with various co-expression structures, our method was very accurate, identifying these structures in 99% of the datasets, except for the overlapping region (79% only). This lower percentage was due to the fact that in some cases the overlapping region was wrongly associated to another co-localised and co-expressed region. This wrong association happened when the correlations between expression level of the genes in the overlapping region were low. In these cases, the values from PCA method were not significant enough resulting to the definition of a unique structure including this overlapping region and the flanking region. When using our method on "real" dataset, we observed that many genomic regions of co-expressed genes were identified, suggesting the existence of large groups of genes transcriptionally linked. These observations are consistent with previously published results (28, 29). Then, our approach provided a way to identify such groups of co-located and co-expressed genes.

Finally, according to our initial hypothesis we compared Hi-C interaction data and co-expression structures defined with our method. As interaction and expression data produced from several and similar samples were not available in public databases, we made the choice to compare different samples, Hi-C data from human lymphoblastoid cell line with a normal karyotype to expression data from 12 representative normal human tissues (26). These expression data were chosen as they represented natural expressions for several human tissues. The method identified numerous regions of co-expressed genes distributed among 4 structures of co-expression. Despite the apparent discrepancy in samples, our comparisons demonstrated that 63% of the genes that were co-expressed according to our method were also found physically interacting using Hi-C method. Furthermore, 88% of the co-expressed regions were found with physical interactions. However, the number of co-expressed regions and common relations were not distributed randomly among the chromosomes and within the structures of co-expression. Some chromosomes (4, 13, 18, and Y for examples) presented few regions of co-expression and common relations, whereas other

chromosomes (1, 17, 19, 22 and X) presented numerous regions and numerous common relations as chromosomes. These results are consistent with previous FISH studies (30–32): chromosomes 4, 13 and 18 are know to tend to be located around the nuclear periphery whereas the chromosomes 1, 17, 19 and 22 are known to be located in the central part of the nucleus. Concerning the distribution of the regions that present physical interaction among the structures, the first structure and the second structure presented 72 % and 20% the regions that present physical interactions, respectively.
We also observed that, 27% of the Hi-C interactions previously published were not associated to co-expression relations. This absence of co-expression for these interactions may be some false positives of Hi-C interactions or real physical interactions with no effect on the gene regulation. Indeed, we could hypothesize that the chromatin proximity does not always involve gene regulation and then co-expression. We also observed 12% of co-expressed regions with no interaction in the Hi-C results, which may be regulated by mechanisms that do not involve a physical interaction. Altogether, these results and especially the percentage of common relations validate the initial hypothesis: gene expressions are in some part a projection of the genome structure. Although it was shown that the chromatin interaction has an impact on the regulation of gene expression, this percentage of co-expressed and physically interacting genes seems very high. At present, we have no formal evidence of a cause-effect relationship between these physical interactions and the co-expression of genes. The mechanisms of regulation of the expression of the concerned genes remain unclear and further analyses will be needed. However, those results pinpoint large groups of genes that interact and are co-expressed, which may ease the identification of common mechanisms of regulation.

.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary Tables 1-2, Supplementary Figures 1-4.

ACKNOWLEDGEMENT

We thank members of our laboratory and Thomas Guyet for stimulating and helpful discussion.

FUNDING

INRA, Agrocampus Ouest; French Ministry in charge of Agriculture (DGER). Funding for open access charge: INRA.
Conflict of interest statement. None declared.

REFERENCES

1. Fraser,P. and Bickmore,W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413-7, 10.1038/nature05916.

2. Babu,M.M., Janga,S.C., de Santiago,I. and Pombo,A. (2008) Eukaryotic gene regulation in three dimensions and its impact on genome evolution. *Current opinion in genetics & development*, **18**, 571-82, 10.1016/j.gde.2008.10.002.

3. Berger,S.L. (2007) The complex language of chromatin regulation during transcription. *Nature*, **447**, 407-12, 10.1038/nature05915.

4. Kharchenko,P.V., Woo,C.J., Tolstorukov,M.Y., Kingston,R.E. and Park,P.J. (2008) Nucleosome positioning in human HOX gene clusters. 1554-1561, 10.1101/gr.075952.107.5.

5. Cook,P.R. (2010) A model for all genomes: the role of transcription factories. *Journal of molecular biology*, **395**, 1-10, 10.1016/j.jmb.2009.10.031.

6. Woodcock,C.L. (2006) Chromatin architecture. *Current opinion in structural biology*, **16**, 213-20, 10.1016/j.sbi.2006.02.005.

7. West,A.G. and Fraser,P. (2005) Remote control of gene transcription. *Human molecular genetics*, **14 Spec No**, R101-11, 10.1093/hmg/ddi104.

8. Göndör,A. and Ohlsson,R. (2009) Chromosome crosstalk in three dimensions. *Nature*, **461**, 212-7, 10.1038/nature08453.

9. Kanduri,C., Pant,V., Loukinov,D., Pugacheva,E., Qi,C.F., Wolffe, a, Ohlsson,R. and Lobanenkov,V.V. (2000) Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Current biology□: CB*, **10**, 853-6.

10. Hark, a T., Schoenherr,C.J., Katz,D.J., Ingram,R.S., Levorse,J.M. and Tilghman,S.M. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, **405**, 486-9, 10.1038/35013106.

11. Bell, a C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482-5, 10.1038/35013100.

12. Court,F., Baniol,M., Hagege,H., Petit,J.S., Lelay-Taha,M.-N., Carbonell,F., Weber,M., Cathala,G. and Forne,T. (2011) Long-range chromatin interactions at the mouse Igf2/H19 locus reveal a novel paternally expressed long non-coding RNA. *Nucleic acids research*, **39**, 5893-906, 10.1093/nar/gkr209.

13. Schneider,R. and Grosschedl,R. (2007) Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes & development*, **21**, 3027-43, 10.1101/gad.1604607.

14. Pombo,A. and Branco,M.R. (2007) Functional organisation of the genome during interphase. *Current opinion in genetics & development*, **17**, 451-5, 10.1016/j.gde.2007.08.008.

15. Misteli,T. (2007) Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787-800, 10.1016/j.cell.2007.01.028.

16. de Laat,W. and Grosveld,F. (2007) Inter-chromosomal gene regulation in the mammalian cell nucleus. *Current opinion in genetics & development*, **17**, 456-64, 10.1016/j.gde.2007.07.009.

17. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science (New York, N.Y.)*, **295**, 1306-11, 10.1126/science.1067799.

18. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., de Wit,E., van Steensel,B. and de Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, **38**, 1348-54, 10.1038/ng1896.

19. Zhao,Z., Tavoosidana,G., Sjölinder,M., Göndör,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Sandhu,K.S., Singh,U., et al. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*, **38**, 1341-7, 10.1038/ng1891.

20. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C., et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. 1299-1309, 10.1101/gr.5571506.1.

21. Li,G., Fullwood,M.J., Xu,H., Mulawadi,F.H., Velkov,S., Vega,V., Ariyaratne,P.N., Mohamed,Y.B., Ooi,H.-S., Tennakoon,C., et al. (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology*, **11**, R22, 10.1186/gb-2010-11-2-r22.

22. Duan,Z., Andronescu,M., Schutz,K., Mcllwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and William,S. (2010) NIH Public Access. **465**, 363-367, 10.1038/nature08973.A.

23. Lieberman-Aiden,E., Van Berkum,N., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B., Sabo,P., Dorschner,M., et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289-293.

24. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K. a, Phillippy,K.H., Sherman,P.M., et al. (2011) NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic acids research*, **39**, D1005-10, 10.1093/nar/gkq1184.

25. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E., et al. (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic acids research*, **39**, D1002-4, 10.1093/nar/gkq1040.

26. Yanai,I., Benjamin,H., Shmoish,M., Chalifa-Caspi,V., Shklar,M., Ophir,R., Bar-Even,A., Horn-Saban,S., Safran,M., Domany,E., et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics (Oxford, England)*, **21**, 650-9, 10.1093/bioinformatics/bti042.

27. Haider,S., Ballester,B., Smedley,D., Zhang,J., Rice,P. and Kasprzyk,A. (2009) BioMart Central Portal--unified access to biological data. *Nucleic acids research*, **37**, W23-7, 10.1093/nar/gkp265.

28. Nie,H., Crooijmans,R.P., Bastiaansen,J.W., Megens,H.J. and Groenen,M.A. (2010) Regional regulation of transcription in the chicken genome. *BMC Genomics*, **11**, 28.

29. Caron,H., Schaik,B., Mee,M., Baas,F., Riggins,G., Sluis,P., Hermus,M., Asperen,R., Boon,K., Voute,P.A., et al. (2001) The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains. *Science*, **291**, 1289-1292.

30. Croft,J. a, Bridger,J.M., Boyle,S., Perry,P., Teague,P. and Bickmore,W. a (1999) Differences in the localization and morphology of chromosomes in the human nucleus. *The Journal of cell biology*, **145**, 1119-31.

31. Boyle,S., Gilchrist,S., Bridger,J.M., Mahy,N.L., Ellis,J. a and Bickmore,W. a (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, **10**, 211-9.

32. Tanabe,H., Habermann,F. a, Solovei,I., Cremer,M. and Cremer,T. (2002) Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutation research*, **504**, 37-45.

33. Miele,A. and Dekker,J. (2008) Long-range chromosomal interactions and gene regulation. *Mol. BioSyst.*, **4**, 1046-1057, 10.1039/B803580F.

TABLE AND FIGURES LEGENDS

Figure 1. FLOWCHARTS OF THE METHOD. The method is based on three dependant steps: first describing the local structure of gene expression by using sequential PCA for co-located genes within

a sliding window; second defining the chromosomal regions where all the window of co-located genes presented the same structure of co-expression by using MFA; and third defining the region of co-expressed genes that presented the same structure of co-expression by using MFA and HAC.

Figure 2. EFFECTS OF THE LENGTH OF THE WINDOW AND THE NUMBER OF CO-EXPRESSED GENES ON THE DESCRIPTION OF LOCAL STRUCTURES OF CO-EXPRESSION.. Datasets of 50 genes with one group of co-expressed genes were simulated. The percentage of true positive gene is the percentage of genes identified among the total co-expressed genes, whereas the false positive percentage is the percentage of false co-expressed genes within the detected region. A: True positive region is the percentage of dataset where a region is identified thanks to the simulated co-expressed genes. The percentage given for true and false positive genes corresponds to the median percentage according to the size of the window. The size of the window does not significantly affect those three results. B: The mean of the percentage of true and false positive genes is given for each size of window according to the number of co-expressed genes simulated for the dataset. The number of co-expressed gene has little effect among detection of true co-expressed genes. Windows of size 15 and 20 minimize the false positive percentage.

Figure 3. EFFECTS OF GENE DENSITY, CORRELATION LEVELS AND NUMBER OF SAMPLES ON THE DESCRIPTION OF LOCAL STRUCTURES OF CO-EXPRESSION. Datasets of 50 genes were simulated where: the density of co-expressed genes inside a genomic region; the level of correlation between these co-expressed genes; and the number of samples conjointly varied. These datasets were analysed with a sliding window of 15 genes (α=0.05). The percentage of true positive gene is the percentage of genes identified among the total co-expressed genes, whereas the false positive percentage is the percentage of false co-expressed genes within the detected region. The median of these percentages is plotted according to each combination 2 from the 3 tested parameters: number of individuals, the level of correlation between the genes and the density of co-expressed genes within a region.

Figure 4. DETECTION OF CO-EXPRESSED REGIONS AND CO-EXPRESSION STRUCTURES **IN SIMULATED DATA.**. Dataset of co-expressed and co-located genes where simulated between two chromosomes of 200 and 210 genes with a level of co-expression of 0.5. Six regions were simulated including: i) one independent region (orange), ii) two interacting regions among the same chromosome (blue), iii) two interacting regions between the two chromosomes (red) and iv) one independent region that overlap two other independent regions (green). chr1_R1= 8 co-expressed genes (among 20 genes); chr1_R2=15(40); chr2_R3=15(45); chr2_R4=25(57); chr2_R5=15(27); and chr2_R6=15(36); The co-expressed genes are plotted according to their location and group of co-expression. 80 % of the results were very similar to the simulated dataset as shown.

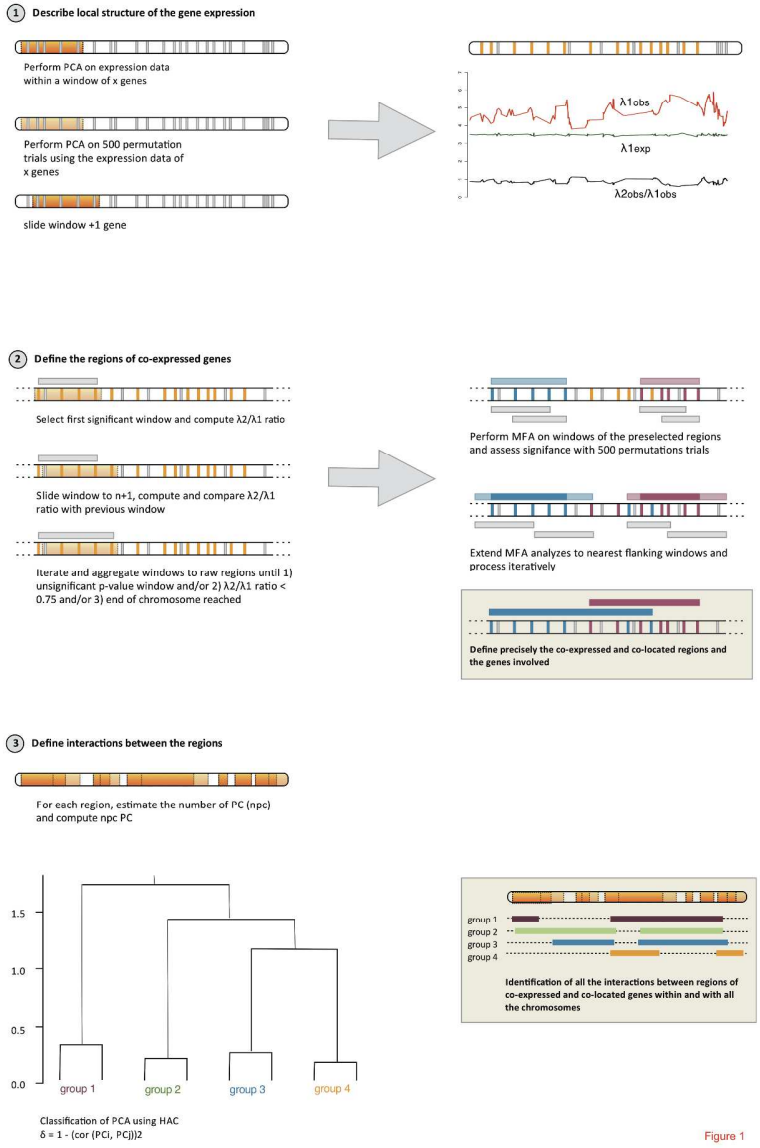
Figure 5. STRUCTURES OF CO_EXPRESSION AND PHYSICAL INTERACTIONS BETWEEN CO-EXPRESSED REGIONS. A: Chromosomes are organized circularly and scaled to their sizes (Mb). The 759 regions of co-expressed genes identified are linked depending one their structure of expression and the physical interactions founded in the Hi-C data. The regions that belong to the same structure and that presented genes found as interacting are linked. Structures of co-expression are shown with different colours: blue for the first structure; green for the second; yellow for the third; and red for the fourth. Relations of co-expression without physical interaction are shown in grey. Intra-chromosomal relations are plotted outside the circle whereas inter-chromosomal relations are plotted inside. The four structures of co-expression are individually represented in B; C; D and E.

SUPPLEMENTARY DATA.

Supplementary Table 1. NUMBER OF CO-EXPRESSION RELATIONS AND INTERACTIONS OF PHYSICAL BETWEEN CO-EXPRESSED REGIONS . The number of specific co-expression relations between the regions (Coexp.) and relations of co-expression found with physical interaction in Hi-C data (Hi-C) are indicated, altogether and for each 4 structures of co-expression and chromosome.

Supplementary Table 2. NUMBER OF CO-EXPRESSED REGIONS AND PHYSICALY INTERACTING REGIONS. The number of regions with specific relations of co-expression between them (Coexp.) and regions with both relations of co-expression and physical interaction (Hi-C) are indicated, altogether and for each 4 structures of co-expression, for each chromosome.

Supplementary Figures. STRUCTURES OF CO_EXPRESSION AND PHYSICAL INTERACTIONS BETWEEN CO-EXPRESSED REGIONS. FOR EACH CHROMOSOME. For each chromosome, the relations of co-expression and the co-expressed regions found with physical interactions in Hi-C data among the genome are represented. Chromosomes are organized circularly and scaled to their sizes (Mb). The relations are defined with different colours depending on the structure of co-expression: blue for the first structure; green for the second; yellow for the third; and red for the fourth. Regions of co-expression without physical interaction are shown in grey. Intra-chromosomal relations are plotted outside the circle whereas inter-chromosomal links are plotted inside. The four structures of co-expressed are also represented individually.



FLOWCHARTS OF THE METHOD. The method is based on three dependant steps: first describing the local structure of gene expression by using sequential PCA for co-located genes within a sliding window; second defining the chromosomal regions where all the window of co-located genes presented the same structure of co-expression by using MFA; and third defining the region of co-expressed genes that presented the same structure of co-expression by using MFA and HAC.

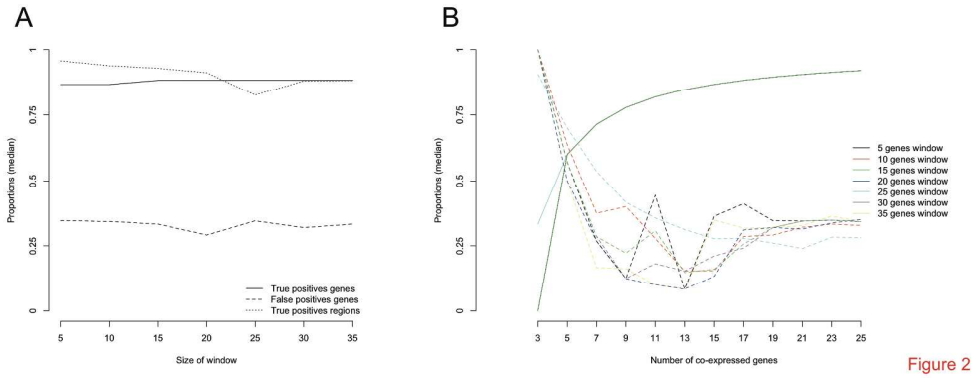


Figure 2. EFFECTS OF THE LENGTH OF THE WINDOW AND THE NUMBER OF CO-EXPRESSED GENES ON THE DESCRIPTION OF LOCAL STRUCTURES OF CO-EXPRESSION.. Datasets of 50 genes with one group of co-expressed genes were simulated. The percentage of true positive gene is the percentage of genes identified among the total co-expressed genes, whereas the false positive percentage is the percentage of false co-expressed genes within the detected region. A: True positive region is the percentage of dataset where a region is identified thanks to the simulated co-expressed genes. The percentage given for true and false positive genes corresponds to the median percentage according to the size of the window. The size of the window does not significantly affect those three results. B: The mean of the percentage of true and false positive genes is given for each size of window according to the number of co-expressed genes simulated for the dataset. The number of co-expressed gene has little effect among detection of true co-expressed genes. Windows of size 15 and 20 minimize the false positive percentage.

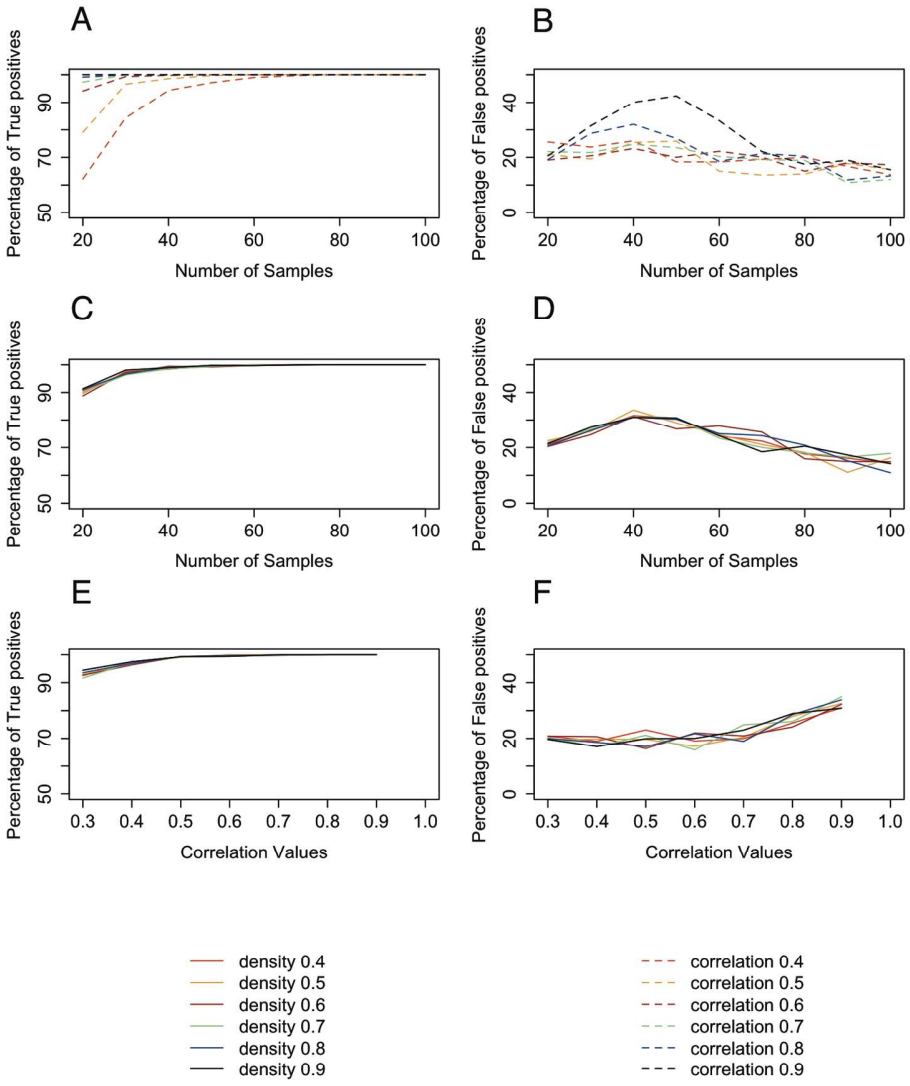


Figure 3

Figure 3. EFFECTS OF GENE DENSITY, CORRELATION LEVELS AND NUMBER OF SAMPLES ON THE DESCRIPTION OF LOCAL STRUCTURES OF CO-EXPRESSION. Datasets of 50 genes were simulated where: the density of co-expressed genes inside a genomic region; the level of correlation between these co-expressed genes; and the number of samples conjointly varied. These datasets were analysed with a sliding window of 15 genes (α=0.05). The percentage of true positive gene is the percentage of genes identified among the total co-expressed genes, whereas the false positive percentage is the percentage of false co-expressed genes within the detected region. The median of these percentages is plotted according to each combination 2 from the 3 tested parameters: number of individuals, the level of correlation between the genes and the density of co-expressed genes within a region.

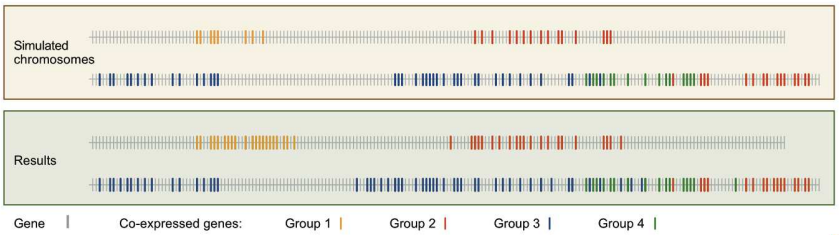


Figure 4

Figure 4. DETECTION OF CO-EXPRESSED REGIONS AND CO-EXPRESSION STRUCTURES IN SIMULATED DATA. Dataset of co-expressed and co-located genes where simulated between two chromosomes of 200 and 210 genes with a level of co-expression of 0.5. Six regions were simulated including: i) one independent region (orange), ii) two interacting regions among the same chromosome (blue), iii) two interacting regions between the two chromosomes (red) and iv) one independent region that overlap two other independent regions (green). chr1_R1= 8 co-expressed genes (among 20 genes); chr1_R2=15(40); chr2_R3=15(45); chr2_R4=25(57); chr2_R5=15(27); and chr2_R6=15(36); The co-expressed genes are plotted according to their location and group of co-expression. 80 % of the results were very similar to the simulated dataset as shown.

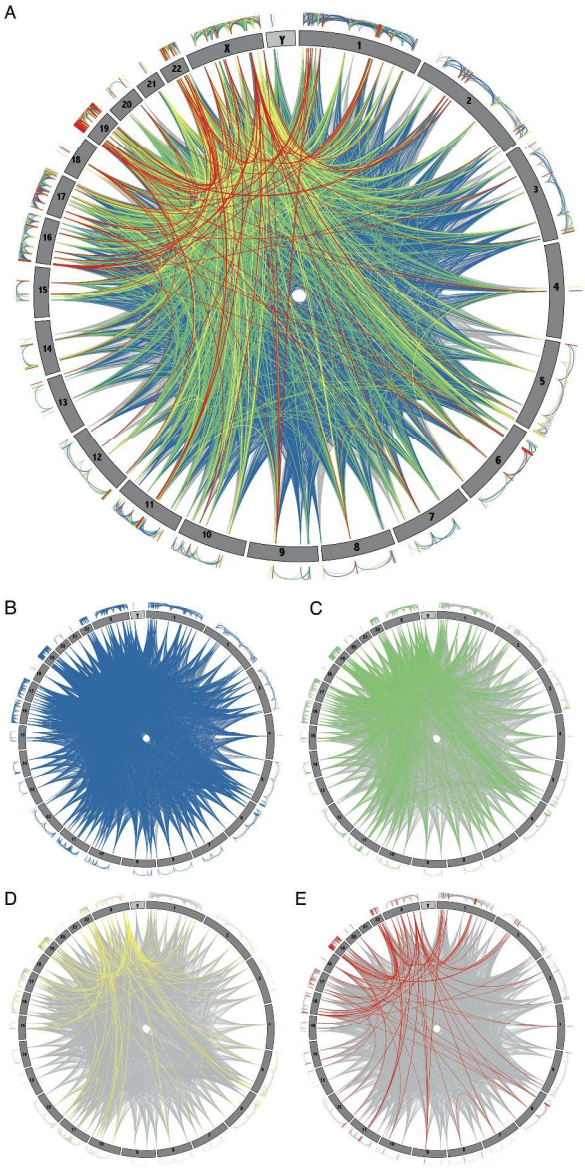


Figure 5. STRUCTURES OF CO_EXPRESSION AND PHYSICAL INTERACTIONS BETWEEN CO-EXPRESSED REGIONS. A: Chromosomes are organized circularly and scaled to their sizes (Mb). The 759 regions of co-expressed genes identified are linked depending one their structure of expression and the physical interactions founded in the Hi-C data. The regions that belong to the same structure and that presented genes found as interacting are linked. Structures of co-expression are shown with different colours: blue for the first structure; green for the second; yellow for the third; and red for the fourth. Relations of co-expression without physical interaction are shown in grey. Intra-chromosomal relations are plotted outside the circle whereas inter-chromosomal relations are plotted inside. The four structures of co-expression are individually represented in B; C; D and E.

CHAPITRE 5

DÉVELOPPEMENTS LOGICIELS ET APPLICATIONS

Dans ce chapitre, nous présentons les développements logiciels associés aux méthodologies mises au point au cours de ce travail de recherche. Chaque méthodologie statistique a été implémentée dans des logiciels libres tels que R ou Scilab. Ainsi, l'ACP régularisée (méthodologie présentée chapitre 2) a été codée à la fois en R et en Scilab (pour faciliter le traitement d'images). L'algorithme de clustering de gènes basé sur l'intégration d'annotations Gene Ontology (méthodologie présentée chapitre 3) a été valorisé sous la forme d'un package `InteGO` pour le logiciel R. Enfin, l'algorithme de prise en compte de la localisation chromosomique (méthodologie présentée chapitre 4), issu d'un travail collaboratif avec le laboratoire de génétique animale d'Agrocampus Ouest, a été implémenté par Marion Ouédraogo sous la forme d'un package `CocoMap` pour le logiciel R. Nous ne détaillerons pas ce dernier.

Sommaire

Au cours de ce travail de recherche, plusieurs travaux de développement logiciel, sur R (R Core Team, 2013) ou Scilab (Scilab Enterprises, 2012) ont été menés ainsi qu'une application externe à l'étude des données transcriptomiques. Le choix du logiciel R est motivé par le fait qu'il est libre et gratuit. De plus R tend à devenir **le** logiciel de statistique de référence puisqu'il est très complet et en essor permanent. Il présente l'avantage certain d'être utilisé à la fois dans le monde universitaire et dans le monde de l'entreprise. Par ailleurs, nous nous sommes intéressés en complément du traitement de données transcriptomiques au traitement d'image dans le développement de l'ACP régularisée. Ainsi, nous avons choisi de développer en Scilab une partie des programmes correspondant au traitement d'images, parce que Scilab est l'équivalent libre du logiciel MATLAB (MATLAB, 2010), référence en traitement d'image.

1 ACP régularisée

Nous proposons donc dans cette section de présenter les deux programmes d'ACP régularisée, objet de l'article Verbanck *et al.* (2013a) : le programme R développé dans l'optique de s'intégrer au package `FactoMineR` (Husson *et al.*, 2013), package dédié à l'analyse exploratoire multivariée, et le programme Scilab tourné vers le traitement d'images. Toutes les fonctions sont disponibles à l'adresse suivante : `http://marie.verbanck.free.fr/packages/rPCA`.

1.1 Programme R d'ACP régularisée

Nous avons donc développé un programme R d'ACP régularisée sous la forme d'une fonction `rPCA()` très simple :

```
rPCA(X, S)
```

Les arguments de cette fonction sont les suivants :

- `X` : une matrice de données quantitative sur laquelle on souhaite faire une ACP régularisée
- `S` : le nombre de dimensions sous-jacentes du signal à prendre en compte dans l'analyse

Cette fonction renvoie une matrice $\hat{\mathbf{X}}$ estimée par ACP régularisée. L'ACP régularisée nécessite un paramètre de réglage qui peut être estimé au préalable par la fonction `estim_ncp()` du package `FactoMineR`.

Concrètement pour réaliser une ACP régularisée sur le jeu de données transcriptomiques *poulets*, qui est présenté dans le chapitre chapitre 1 section 3, voici comment il faut procéder :

```
> Xhat = rPCA(X = poulets, S = 3)
```

Nous prenons en compte trois dimensions sous-jacentes car les poulets du jeu de données ont été soumis à quatre conditions expérimentales, nous nous attendons par conséquent à avoir un signal en trois dimensions. La fonction `rPCA()` fournit la matrice *Xhat* qui est l'estimation du jeu de données *poulets* par ACP régularisée. Pour obtenir ensuite les représentations graphiques associées à cette ACP régularisée, il faut utiliser la fonction `PCA()` de `FactoMineR` avec l'argument `scale.unit` qui doit impérativement être égal à `FALSE`. Nous ajoutons également la variable de conditions expérimentales (*conditionsExperimentales*) à la matrice *Xhat* afin de l'ajouter en tant que variable qualitative supplémentaire (figure 1) :

```
> Xhat.rpca = PCA(cbind.data.frame(conditionsExperimentales,
    Xhat), scale.unit = FALSE, ncp = 3, quali.sup = 1, graph =
    TRUE, axes = c(1,2))
```

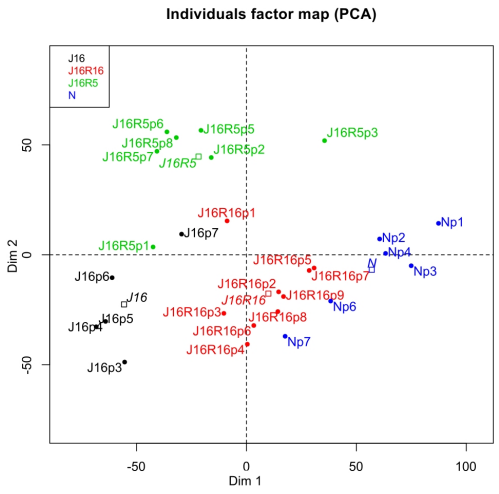


FIGURE 1 – Premier plan factoriel de l'ACP régularisée du jeu de données *poulets*.

Enfin, la matrice $\hat{\mathbf{X}}$ estimée par ACP régularisée peut être utilisée comme telle dans une autre procédure, un Heatmap par exemple (figure 2) :

```
> heatmap.rpca = heatmap(Xhat, main = 'rPCA')
```

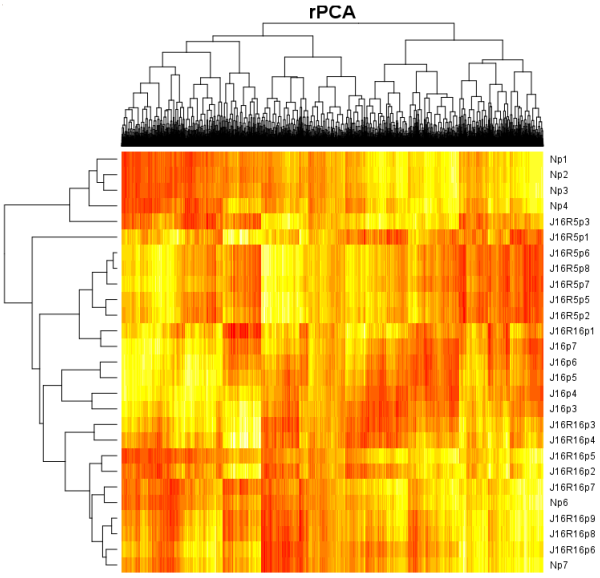


FIGURE 2 – Heatmap suite à l'ACP régularisée du jeu de données *poulets*.

1.2 Programme Scilab d'ACP régularisée à travers le traitement du jeu de données PINCAT

Nous avons également développé un programme Scilab d'ACP régularisée. Nous avons fait ce choix principalement pour pouvoir traiter des images qui peuvent être codées sous forme de matrices complexes. Sachant que le logiciel R n'est pas le mieux adapté pour traiter les matrices complexes nous avons choisi de nous tourner vers un logiciel libre adapté Scilab. Nous présentons le programme développé à partir d'un exemple de données d'image, le jeu de données PINCAT (Sharif et Bresler, 2007) qui est utilisé dans Candès *et al.* (2012) et que nous avons présenté dans le chapitre 1 section 3. Ce programme permet notamment de réaliser le plan de simulations proposé dans l'article Verbanck *et al.* (2013a) à partir du jeu de données PINCAT.

```
rPCA_images(image, S, sigma, lambda, time)
```

Les arguments de cette fonction sont les suivants :

- `image` : cube de données avec sur les deux premières dimensions les images à proprement parler, avec à l'intersection d'une ligne et d'une colonne un pixel, et dans la profondeur les différentes images (qui correspondent à des temps ici).
- `S` : le nombre de dimensions sous-jacentes du signal à prendre en compte dans l'analyse

- `sigma` : l'écart-type du bruit
- `lambda` : valeur de la correction SURE (Candès *et al.*, 2012)
- `time` : numéro de l'image utilisée pour les sorties graphiques

Cette fonction renvoie un vecteur avec les trois valeurs d'erreur quadratique moyenne, la première étant l'erreur associée au débruitage par ACP, la seconde par ACP régularisée et la troisième par la méthode SURE. De plus elle renvoie un graphique composé de cinq images avec de gauche à droite, la vraie image, c'est-à-dire l'image correspondant au signal, l'image bruitée, l'image débruitée par ACP, l'image débruitée par ACP régularisée et l'image débruitée par la méthode SURE (figure 3).

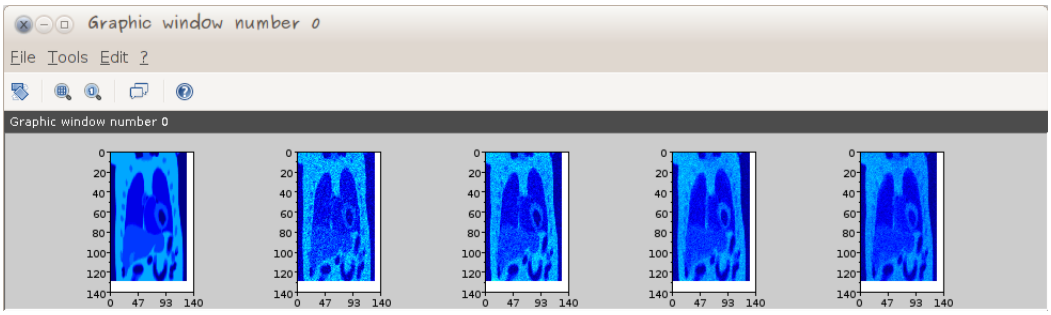


FIGURE 3 – Sortie graphique de la fonction Scilab `rPCA_images()`, la fonction est appliquée aux données PINCAT. Il s'agit ici de l'image au temps numéro 5 qui est représentée.

2 PRÉSENTATION DU PACKAGE INTEGO

Le package `InteGO` rassemble toutes les fonctionnalités qui ont été développées dans le cadre de l'intégration d'information Gene Ontology dans l'article Verbanck *et al.* (2013b). Ce package fournit et évalue des clusters de gènes en intégrant, aux données d'expression, de l'information extérieure sur les gènes, sous la forme d'annotations fonctionnelles Gene Ontology. `InteGO` est téléchargeable à l'adresse suivante : `http://marie.verbanck.free.fr/packages/InteGO`.

2.1 DONNÉES

Concrètement, pour utiliser le package, il faut disposer de deux jeux de données : un *jeu de données transcriptomiques* et un *jeu de données d'annotations*.

Pour être utilisable dans le package, le *jeu de données transcriptomiques* doit être sous la forme suivante : il doit s'agir d'un jeu de données quantitatif avec en ligne les gènes et en colonne les sujets. Le package tolère la présence de données manquantes dans le jeu de données. Une donnée d'expression manquante pour un gène sera alors imputée par la moyenne des expressions du gène en question en excluant les données manquantes. Ainsi, si l'utilisateur préfère utiliser une autre technique d'imputation des données

manquantes (par exemple Josse *et al.* (2011)), il est préférable de réaliser l'imputation au préalable afin d'obtenir un jeu de données complété qui pourra être utilisé dans `InteGO`.

Le *jeu de données d'annotations* doit comporter des annotations dans lesquelles on associe des éléments à chaque gène. Par exemple, dans les annotations de type Gene Ontology, à chaque gène est associé une liste de termes. Pour être utilisable dans le package, le jeu de données d'annotations doit être sous la forme suivante : il doit s'agir d'un jeu de données quantitatif avec en ligne les gènes, en colonne toutes les annotations (une colonne correspond par exemple à un terme Gene Ontology). À l'intersection d'une ligne et d'une colonne du jeu de données, se trouve 1 si le gène est associé à l'annotation en question, 0 sinon.

Pour présenter le package, nous proposons, dans un premier temps, d'exposer succinctement les fonctions permettant de mettre en œuvre chaque étape de l'algorithme : l'intégration d'information biologique, le clustering puis l'évaluation des clusters. Ensuite, nous présenterons la fonction principale du package, qui permet de réaliser toutes les étapes de l'algorithme, avec un exemple d'interprétation.

2.2 Présentation des étapes de l'algorithme

2.2.1 Intégration d'information biologique : obtention de fonctions biologiques coexprimées

La première étape de l'algorithme consiste à combiner l'information biologique aux données d'expression, il s'agit de l'obtention de fonctions biologiques coexprimées. Pour construire les fonctions biologiques coexprimées, il faut utiliser la fonction `Integration()` de la façon suivante :

```
Integration(annotations, expressions, nb.dim.ex, LIM.ASSO =
    10, LIM.COR = 0.6)
```

Les arguments de cette fonction sont les suivants :

- `annotations` : jeu de données d'annotations, par exemple Gene Ontology. Les données doivent être codées de la façon suivante : les gènes en ligne et les annotations en colonne. Chaque colonne correspond donc à une annotation, par exemple un terme Gene Ontology. À l'intersection d'une ligne et d'une colonne du jeu de données, se trouve 1 si le gène est associé à l'annotation en question, 0 sinon. Ce tableau de données est quantitatif.
- `expressions` : jeu de données transcriptomiques avec en ligne les gènes et en colonne les sujets. Ce jeu de données est quantitatif et tolère la présence de données manquantes qui sont estimées par la moyenne.
- `nb.dim.ex` : nombre de dimensions sous-jacentes du jeu de données transcriptomiques. Ce nombre est important car seules les `nb.dim.ex` premières dimensions seront prises en compte dans la procédure de clustering. Si cet argument est `NULL` (par défaut), toutes les dimensions sont prises en compte.

- `LIM.ASSO` : seuil en dessous duquel une fonction biologique n'est pas décomposée en plusieurs fonctions biologiques coexprimées. Ce seuil correspond à un nombre de gènes associés à une fonction biologique.
- `LIM.COR` : seuil au dessus duquel une fonction biologique peut être considérée comme une fonction biologique coexprimée et ne sera pas décomposée en plusieurs fonctions biologiques coexprimées. Ce seuil correspond une valeur de l'indicateur de coexpression comprise entre 0 et 1.

Cette fonction retourne :

- `annotations.sep` : une matrice binaire équivalent à la matrice `annotations` mais qui code les associations entre les gènes et les fonctions biologiques coexprimées.

2.2.2 Obtention de clusters de gènes

Une fois le tableau croisant gènes et fonctions biologiques coexprimées obtenu, nous pouvons appliquer l'algorithme de clustering afin d'obtenir les clusters de gènes. Pour mettre en œuvre l'algorithme de clustering, il faut utiliser la fonction `GeneClustering()` :

```
GeneClustering(annotations.sep, nb.dim.an = NULL, nb.group)
```

- `annotations.sep` : une matrice binaire équivalent à la matrice `annotations` mais qui code les associations entre les gènes et les fonctions biologiques coexprimées. Les données sont donc codées de la façon suivante : les gènes en ligne et les fonctions biologiques coexprimées en colonne. À l'intersection d'une ligne et d'une colonne du jeu de données, se trouve 1 si le gène est associé à la fonction biologique coexprimée en question, 0 sinon. Ce tableau de données est quantitatif. La matrice annotations.sep est retournée par la fonction `Integration()`.
- `nb.dim.an` : nombre de dimensions sous-jacentes du jeu de données d'annotations. Ce nombre est important car seules les `nb.dim.an` premières dimensions seront prises en compte dans la procédure de clustering. Si cet argument est `NULL` (par défaut), toutes les dimensions sont prises en compte.
- `nb.group` : nombre de clusters de gènes à construire

Cette fonction renvoie une liste :

- `groups` : liste dans laquelle chaque élément représente un cluster de gènes, autrement dit chaque élément est un vecteur contenant les identifiants des gènes du cluster.

Une fois les clusters de gènes obtenus, nous pouvons évaluer leur caractéristiques en termes de coexpression et d'homogénéité biologique.

2.2.3 Évaluation des clusters de gènes

L'évaluation des clusters de gènes s'effectue au moyen de deux fonctions, la fonction `Indicators()` qui permet de calculer les valeurs des indicateurs de coexpression et d'homogénéité biologique et la fonction `IndicatorsPvalues()` qui permet d'associer à chaque

indicateur une probabilité critique. Intéressons-nous d'abord à la fonction `Indicators()` qui s'utilise ainsi :

```
Indicators(groups, expressions, annotations)
```

Les arguments de cette fonction sont :

- `groups` : liste dans laquelle chaque élément représente un cluster de gènes : chaque élément est un vecteur contenant les identifiants des gènes du cluster. Cette liste est retournée par la fonction `GeneClustering()`.
- `expressions` : jeu de données transcriptomiques quantitatif avec en ligne les gènes et en colonne les sujets.
- `annotations` : jeu de données d'annotations quantitatif avec les gènes en ligne et les annotations en colonne. À l'intersection d'une ligne et d'une colonne, se trouve 1 si le gène est associé à l'annotation en question, 0 sinon.

Cette fonction retourne :

- `groups.indic` : liste dans laquelle chaque élément correspond à un cluster. Chaque élément est un vecteur composé de deux valeurs pour le cluster en question, la première correspondant à l'indicateur de coexpression et la seconde à l'indicateur d'homogénéité biologique.

Ensuite, la fonction `IndicatorsPvalues()` permet de calculer les probabilités critiques associées. Elle s'utilise ainsi :

```
IndicatorsPvalues(abaque = NULL, groups, groups.indic,
   expressions, annotations, NB.SIM = 100)
```

Les arguments de cette fonction sont :

- `abaque` : cet argument est optionnel et est une astuce pour gagner du temps sur l'estimation des distributions de probabilités associées aux deux indicateurs, indicateur de coexpression et indicateur d'homogénéité biologique. Il s'agit d'une liste contenant les distributions déjà obtenues sur les mêmes données.
- `groups` : liste dans laquelle chaque élément représente un cluster de gènes : chaque élément est un vecteur contenant les identifiants des gènes du cluster. Cette liste est retournée par la fonction `GeneClustering()`.
- `groups.indic` : liste dans laquelle chaque élément correspond à un cluster : chaque élément est un vecteur composé de deux valeurs pour le cluster en question, la première correspondant à l'indicateur de coexpression et la seconde à l'indicateur d'homogénéité biologique. Cette liste est retournée par la fonction `Indicators()`.
- `expressions` : jeu de données transcriptomiques quantitatif avec en ligne les gènes et en colonne les sujets.
- `annotations` : jeu de données d'annotations quantitatif avec les gènes en ligne et les annotations en colonne. À l'intersection d'une ligne et d'une colonne, se trouve 1 si le gène est associé à l'annotation en question, 0 sinon.
- `NB.SIM` : nombre de simulations à réaliser pour estimer les distributions de probabilités associées aux deux indicateurs. Par défaut cet argument est fixé à 100.

Cette fonction retourne :

- `groups.pvalues` : liste dans laquelle chaque élément correspond à un cluster. Chaque élément est un vecteur composé de deux valeurs pour le cluster en question, la première correspondant à la probabilité critique de l'indicateur de coexpression et la seconde à la probabilité critique de l'indicateur d'homogénéité biologique.

2.3 Fonction principale : `intego()`

Toutes les étapes que nous venons de présenter sont rassemblées dans une seule et même fonction, la fonction `intego()`. Cette fonction permet à partir d'un jeu de données d'expression et d'un jeu de données d'annotations, d'obtenir une partition des gènes puis d'évaluer les clusters de gènes obtenus par la procédure proposée. Elle s'utilise de la façon suivante :

```
intego(expressions, nb.dim.ex = NULL, annotations, nb.dim.an
    = NULL, nb.group, abaque = NULL, NB.SIM = 100, WRITE =
   FALSE, GRAPH = TRUE, LIM.ASSO = 4, LIM.COR = 0.5)
```

- `expressions` : jeu de données transcriptomiques quantitatif avec en ligne les gènes et en colonne les sujets.
- `nb.dim.ex` : nombre de dimensions sous-jacentes du jeu de données transcriptomiques
- `annotations` : jeu de données d'annotations quantitatif avec les gènes en ligne et les annotations en colonne. À l'intersection d'une ligne et d'une colonne, se trouve 1 si le gène est associé à l'annotation en question, 0 sinon.
- `nb.dim.an` : nombre de dimensions sous-jacentes du jeu de données d'annotations.
- `nb.group` : nombre de clusters de gènes à construire.
- `abaque` : argument optionnel. Liste contenant les distributions des indicateurs déjà obtenues sur les mêmes données.
- `NB.SIM` : nombre de simulations à réaliser pour estimer les distributions de probabilités associées aux deux indicateurs. Par défaut cet argument est fixé à 100.
- `WRITE` : argument booléen, s'il est vrai l'abaque contenant toutes les distributions de probabilités associées aux indicateurs est écrit dans un fichier du répertoire courant. Par défaut le fichier n'est pas écrit.
- `GRAPH` : argument booléen, s'il est vrai le graphique représentant les probabilités critiques est produit. Par défaut le graphique est produit.
- `LIM.ASSO` : seuil en dessous duquel une fonction biologique n'est pas décomposée en plusieurs fonctions biologiques coexprimées. Ce seuil correspond à un nombre de gènes associés à une fonction biologique.
- `LIM.COR` : seuil au dessus duquel une fonction biologique peut être considérée comme une fonction biologique coexprimée et ne sera pas décomposée en plusieurs fonctions biologiques coexprimées. Ce seuil correspond une valeur de l'indicateur de coexpression comprise entre 0 et 1.

La fonction retourne les éléments suivants :

- `groups` : liste dans laquelle chaque élément représente un cluster de gènes, autrement dit chaque élément est un vecteur contenant les identifiants des gènes du cluster.
- `indicators` : liste dans laquelle chaque élément contient les deux valeurs des indicateurs de coexpression et d'homogénéité biologique, pour un cluster de gènes.
- `pvalues` : liste dans laquelle chaque élément contient les deux probabilités critiques associées aux indicateurs de coexpression et d'homogénéité biologique, pour un cluster de gènes.
- `abaque` : liste contenant les distributions utilisées pour calculer les probabilités critiques associées avec les deux indicateurs.

Pour comprendre concrètement comment utiliser la fonction `intego()`, nous proposons de l'appliquer au jeu de données *poulets* associé à un jeu de données d'annotations *GO* :

```
> res.intego <- intego(expressions = poulets, nb.dim.ex = 3,
    annotations = GO, nb.dim.an = NULL, nb.group = 10, abaque =
     NULL, NB.SIM = 100, GRAPH = TRUE)
```

La fonction renvoie des clusters de gènes décrits dans l'objet `groups` :

```
> res.intego$groups
$Group.1
[1] "RIGG18189" "RIGG03060" "RIGG05375" "RIGG20090" "RIGG18949
   " "RIGG00124"
[7] "RIGG12020" "RIGG03740"
$Group.2
 [1] "RIGG06180" "RIGG12025" "RIGG19070" "RIGG18185" "
    RIGG13105" "RIGG13362"
 [7] "RIGG17069" "RIGG05835" "RIGG06582" "RIGG07906" "
    RIGG16744" "RIGG01014"
[13] "RIGG13051" "RIGG11729" "RIGG05774" "RIGG15257" "
   RIGG00767" "RIGG05924"
...
```

Il s'agit d'une liste dont chaque élément fournit les identifiants des gènes appartenant à un cluster. Une fois les clusters obtenus, nous pouvons nous intéresser à leurs caractéristiques à travers les probabilités critiques associées aux deux indicateurs. Les probabilités critiques sont fournies dans l'objet `pvalues` où `indic.coexp` représente la probabilité critique associée à l'indicateur de coexpression tandis que `indic.bh` représente la probabilité critique associée à l'indicateur d'homogénéité biologique :

```
> res.intego$pvalues
$Group.1
indic.coexp     indic.bh
```

```
       0.00         0.05
$Group.2
indic.coexp     indic.bh
       0.81         0.00
...
```

Les probabitlités critiques sont également représentées sous forme graphique si l'argument GRAPH de la fonction intego() est vrai (figure 4), ce qui permet d'avoir une idée générale de la qualité des clusters.

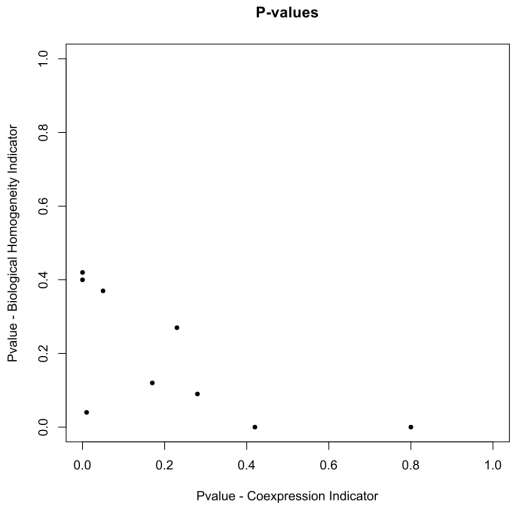


FIGURE 4 – Représentation des probabilités critiques associées à l'indicateur de coexpression (en abscisse) et à l'indicateur d'homogénéité biologique (en ordonnée).

Parmi les clusters obtenus, nous pouvons sélectionner les bons candidats à l'interprétation qui sont les clusters rassemblant des gènes significativement coexprimés et significativement biologiquement apparentés. Nous pouvons associer à ces bons candidats des termes Gene Ontology spécifiques au moyen de tests d'enrichissement.

Par ailleurs, nous souhaitons apporter quelques précisions quant à l'utilisation de l'élément abaque. En effet, nous avons déjà précisé qu'il s'agit d'une astuce permettant de gagner du temps sur l'estimation des distributions de probabilités associées aux deux indicateurs. En effet, le calcul des distributions permettant d'estimer les probabilités critiques est l'étape la plus longue. C'est pourquoi, nous laissons la possibilité d'exécuter cette étape en plusieurs fois. Concrètement dans l'exemple que nous venons de présenter, nous avons estimé les probabilités avec des distributions basées sur 100 simulations, ce qui est spécifié par l'argument NB.SIM. Nous pouvons relancer ce calcul en choisissant de calculer les probabilités avec des distributions basées sur 300 simulations, mais en spécifiant dans l'argument abaque, l'abaque déjà calculée sur 100 simulations, à savoir abaque = res.intego$abaque, la fonction rajoute uniquement 200 simulations aux 100 déjà effectuées :

```
> res.intego1 <- intego(expressions = poulets, nb.dim.ex = 3,
    annotations = GO, nb.dim.an = NULL, nb.group = 10, abaque =
     res.intego$abaque, NB.SIM = 300, GRAPH = TRUE)
```

Cela affine le calcul des probabilités critiques qui sont légèrement modifiées :

```
> res.intego1$pvalues
$Group.1
indic.coexp     indic.bh
       0.02         0.06

$Group.2
indic.coexp     indic.bh
  0.7966667    0.0000000
...
```

2.4 Plan de simulations

Dans l'article Verbanck *et al.* (2013b), nous proposons un plan de simulation des données d'expression et des données d'annotations qui nous a permis de valider notre méthode. Nous n'avons pas inclus, dans le package `InteGO`, les fonctions permettant de simuler ces deux types de données mais les deux scripts correspondants, simulation.exp.R et simulation.go.R, sont accessibles à `http://marie.verbanck.free.fr/packages/InteGO`. L'idée de ce plan de simulations est de simuler premièrement des données d'expression. Puis nous simulons des données d'annotations à partir des données d'expressions sachant qu'une partie des annotations a une structure similaire aux données d'expression tandis que l'autre partie des annotations est aléatoire. Nous proposons de présenter les deux fonctions.

Tout d'abord la fonction `simulation.exp()` permet de simuler des données d'expression. Elle s'utilise de la façon suivante :

```
simulation.exp(n , p, S)
```

avec

- `n` : nombre entier représentant le nombre de sujets du jeu de données d'expression simulé.
- `p` : nombre entier représentant le nombre de gènes du jeu de données d'expression simulé.
- `S` : nombre entier représentant le nombre de dimensions sous-jacentes du jeu de données d'expression simulé.

Ensuite, pour simuler des données d'annotations à partir des données d'expression simulées, il faut utiliser la fonction `simulation.ann()` de la façon suivante :

```
simulation.ann(expressions, ALEA = 2)
```

avec

- `expressions` : matrice quantitative de dimension nombre de gènes $\times$ nombre de sujets. Jeu de données d'expression à partir duquel on construit le jeu de données d'annotations simulé. Il peut s'agit du jeu de données d'expression simulé fourni par la fonction `simulation.exp()`.
- `ALEA` : entier représentant le degré d'annotations aléatoires dans le jeu de données d'annotations simulé. Il s'agit du rapport entre annotations aléatoires et annotations similaires à l'expression : si `ALEA` = 2 (valeur par défaut), il y a deux fois plus d'annotations aléatoires que d'annotations liées au jeu de données `expressions`.

À partir de ces deux jeux de données simulés, nous pouvons utiliser les fonctions du package `InteGO`.

CHAPITRE 6

CONCLUSION ET PERSPECTIVES

L'apparition des données générées à haut débit, telles que les données transcriptomiques (puce à ADN), a offert des opportunités sans précédent de comprendre les mécanismes cellulaires. Ce travail de recherche apporte une contribution à l'analyse de données transcriptomiques. Il a été positionné par rapport à la stratégie classique d'analyse des données transcriptomiques afin d'en faire ressortir les nouvelles problématiques soulevées dans ce travail ainsi que les méthodologies proposées. Ce travail de recherche présente à la fois une unité dans les applications qui sont centrées autour de l'analyse de données transcriptomiques et dans les méthodologies développées qui sont axées sur l'analyse multidimensionnelle exploratoire.

Le chapitre 1 de ce manuscrit présente le contexte de ce travail de recherche. Ce premier chapitre est certes introductif mais il a véritablement constitué une grande partie du travail de recherche. En effet, dans un contexte très appliqué, tous les aspects de traduction entre le domaine d'application et les statistiques ne sont pas à négliger, particulièrement dans un contexte d'étude biologique. Ainsi, connaître et prendre en compte la nature fine des données est crucial dans le développement de méthodologies adaptées. Dans ce premier chapitre, nous proposons donc des rappels et des généralités à la fois biologiques et statistiques. Puis nous introduisons les méthodologies développées en conservant les allers-retours entre biologie et statistique afin de montrer l'adéquation des méthodologies développées avec le domaine d'application.

Le chapitre 2 est centré autour du développement de l'ACP régularisée. Après avoir rappelé quelques considérations générales autour de l'application de l'ACP aux données transcriptomiques, nous proposons un point de vue « modèle » sur l'ACP à travers le modèle à effet fixe de l'ACP (Caussinus, 1986). Ce nouveau point de vue nous permet de mettre en évidence la fonction de débruitage de l'ACP en complément de son rôle dans la visualisation des données. Ainsi, lorsque les données peuvent être vues comme un mélange de signal que l'on cherche à retrouver mais que l'on ne connaît pas, et de bruit,

l'ACP permet de donner une estimation du signal sous-jacent. C'est dans ce cadre que nous avons proposé une version régularisée de l'ACP qui permet de fournir une meilleure estimation (au sens de l'erreur quadratique moyenne) du signal sous-jacent.

L'originalité de notre approche est de proposer une version de l'ACP régularisée avec un terme qui se calcule explicitement à partir des données et qui ne recourt donc pas à la validation croisée très coûteuse en temps de calcul. L'ACP régularisée a ainsi été appliquée avec bénéfices aux données transcriptomiques en permettant d'améliorer la visualisation de ces données ainsi que le pré-traitement à une étape de clustering (Heatmap). De plus, nous avons utilisé l'ACP régularisée en tant qu'outil de traitement d'images. Dans ce cadre, nous avons montré que l'ACP régularisée est un outil très prometteur pour débruiter les images.

Il reste cependant un paramètre de réglage à l'ACP régularisée qui est le nombre de dimensions sous-jacentes du signal. Il existe, dans la littérature, des algorithmes qui permettent d'estimer ce nombre de dimensions, nous pouvons citer l'exemple de Josse et Husson (2011) qui donne des résultats satisfaisants. Une des perspectives de ce travail consiste à régler ce problème d'estimation du nombre de dimensions sous-jacentes. Nous pourrions en effet proposer une estimation du nombre de dimensions sous-jacentes en minimisant un critère qui approche l'erreur de prédiction comme dans Josse et Husson (2011). Cependant, le calcul de ce critère pourrait être étendu aux données reconstituées par ACP régularisée et non pas par ACP.

Par ailleurs, nous envisageons d'étendre la régularisation à l'analyse factorielle multiple (AFM) qui permet de prendre en compte une structure en groupes sur les variables. Sachant que l'AFM est une ACP pondérée, l'extension est naturelle. Nous pourrions utiliser l'AFM régularisée dans l'analyse simultanée de plusieurs tableaux de données *omiques*, tels que des données transcriptomiques couplées à des données protéomiques collectées chez les mêmes sujets par exemple.

Le chapitre 3 est focalisé sur l'intégration d'information biologique de type Gene Ontology dans l'analyse de données d'expression. Nous proposons ainsi un nouvel algorithme de clustering de gènes basé sur une distance entre gènes qui combine les deux informations. Selon cette distance, deux gènes sont proches s'ils présentent à la fois des signatures fonctionnelles (association à des termes Gene Ontology) et des signatures d'expression similaires. Les clusters ainsi obtenus sont ensuite évalués au moyen d'une procédure que nous avons également mise au point. Cette dernière est basée sur deux indicateurs, un indicateur de coexpression et un indicateur d'homogénéité biologique, qui sont associés à une probabilité critique.

L'originalité de notre approche est, d'une part, de prendre en compte, de façon active, l'information biologique, afin de la combiner aux données d'expression. D'autre part, notre méthodologie est originale car elle fournit une procédure d'évaluation objective des clusters de gènes. Les résultats sont très encourageants car notre algorithme fournit une grande proportion de clusters qui sont de bons candidats à l'interprétation puisqu'ils rassemblent des gènes à la fois coexprimés et biologiquement apparentés.

Comme nous l'avons exposé chapitre 3, dans la nomenclature Gene Ontology, les as-

sociations des gènes aux termes sont décrétées d'après plusieurs sources (vérification expérimentale, comparaison in silico, déduction des deux précédentes, source inconnue) dont nous n'avons pas tenu compte pour des raisons de simplification. Cependant, l'extension des méthodologies développées à la prise en compte de ces différentes sources est une possibilité. Nous pourrions imaginer coder les associations entre gènes et GO termes de la façon suivante : 5 modalités représentant les 4 grands types de sources et une modalité représentant la non association. Nous pourrions également imaginer prendre en compte plus finement la fiabilité de la source de façon quantitative. Par ailleurs, il serait souhaitable de relier le package `InteGO` avec la base de données Gene Ontology afin de récupérer directement les annotations GO à partir d'une liste de gènes. Il serait envisagable d'établir une dépendance entre `InteGO` et le package `goTools` (Yang et Paquet, 2009) par exemple.

Dans le chapitre 4, nous présentons un travail collaboratif centré autour du développement d'une méthodologie d'analyse des données transcriptomiques en lien avec la localisation chromosomique. La méthodologie que nous proposons se compose de trois étapes. Premièrement, nous étudions, au moyen d'une fenêtre glissante le long de chaque chromosome, la coexpression des gènes de la fenêtre. Cette fenêtre nous permet ainsi d'identifier des clusters de gènes colocalisés (car contigus sur le chromosome) et coexprimés que nous appelons régions. Une fois les régions chromosomiques primitives définies, la deuxième étape de l'algorithme consiste à affiner ces régions et à détecter la présence de sous-régions qui peuvent se chevaucher. Enfin, une fois toutes les régions et sous-régions définies, la troisième étape consiste à comparer les régions (voire les sous-régions) entre elles afin d'identifier des régions qui peuvent être proches dans l'espace. Les régions ainsi identifiées, ainsi que les regroupements de régions définis sont ensuite confrontés à des données de cartographie d'interactions physiques entre chromosomes (Hi-C).

L'originalité de notre approche réside dans la prise en compte de la localisation des gènes. En effet, les questionnements autour de l'architecture du génome et le positionnement des chromosomes dans le noyau sont très récents et en plein développement. Les résultats sont très prometteurs et la comparaison des régions identifiées par notre méthode avec les interactions physiques montre de grandes similitudes, ce qui est extrêmement encourageant.

Il reste cependant quelques choix techniques à discuter et donc à améliorer. Nous réalisons notamment plusieurs tests d'hypothèse basés sur des techniques d'échantillonnage qui pourraient être améliorées.

En conclusion, ce travail de recherche en statistique appliquée a été l'occasion de se focaliser sur des aspects théoriques (chapitre 2) et sur des aspects très appliqués (chapitres 3 et 4) qui se complémentent bien. De plus, ce travail a été l'occasion d'une collaboration très enrichissante avec des généticiens qui est très formatrice dans un contexte appliqué. Par ailleurs, nous avons pu à la fois proposer des améliorations pour des stratégies d'analyse classique et de nouveaux points de vue sur les données transcriptomiques à tra-

vers les différentes méthodologies d'intégration d'information.

Plus généralement, rappelons que les données transcriptomiques de type puce à ADN ont été l'objet de ce travail de recherche. Cependant, nous savons que ces données présentent certaines limitations et que les données dites de RNA-Seq, qui consistent à séquencer les ARNm sont en plein essor. Néanmoins les traitements de ces données RNA-Seq sont encore balbutiants. Les recueils de données de puce à ADN ont l'avantage d'être très nombreux et beaucoup de ces jeux de données sont publics, ce qui facilite les développements méthodologiques. Nous pouvons cependant envisager d'étendre toutes les méthodologies développées à l'étude des données RNA-Seq.

Enfin, ce travail de recherche a été pleinement tourné vers le développement de méthodologies statistiques, ce qui explique que les interprétations biologiques n'aient pas été poussées. Nous en sommes conscients mais ce doctorat a été l'occasion de se focaliser sur les statistiques avec un objectif de formation assumé. Une interprétation biologique poussée nécessaire à la validation de toutes les méthodes développées reste donc à faire. Mais je souhaite par la suite mettre au service des biologistes toutes les compétences que j'ai pu acquérir pendant ce travail tourné vers la méthodologie et appliquer les méthodologies développées dans le cadre de problématiques biologiques spécifiques, comme j'ai déjà pu le faire au cours de la collaboration avec le laboratoire de génétique animale d'Agrocampus Ouest.

CHAPITRE 7

LISTE DES TRAVAUX

PUBLICATIONS

Marion Ouédraogo, Sébastien Lê, **Marie Verbanck**, Christian Diot, and Frédéric Lecerf, *Identification of gene co-expression structures by multivariate analyses of expression data and comparison with genome interactions*, Nucleic Acid Research (soumis) (2013).

Marie Verbanck, Julie Josse, and François Husson, *Regularised PCA to denoise and visualise data*, Statistics and Computing (soumis) (2013).

Marie Verbanck, Sébastien Lê, and Jérôme Pagès, *A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data*, BMC Bioinformatics **14** (2013), no. 1, 42 (en), Highly Accessed.

COMMUNICATIONS ORALES

Le nom de l'orateur est en premier

Marie Verbanck, Sébastien Lê, and Jérôme Pagès. Integrating biological knowledge related to coexpression when analysing xomic data. Paris, France, August 22-27 2010. 19th International Conference on Computational Statistics.

Marie Verbanck, Sébastien Lê, and Jérôme Pagès. Towards the integration of external information combining biological knowledge and coexpression when interpreting xomic data in an exploratory framework. Paris, France, January 27-28 2011. Statistical Methods for Post Genomic Data (SMPGD).

Marie Verbanck, Sébastien Lê, and Jérôme Pagès. Towards the integration of biological knowledge with canonical correspondence analysis when analyzing xomic data in an

exploratory framework. Rennes, France, February 8-11 2011. Correspondance Analysis and Related MEthods (CARME).

Marie Verbanck, Sébastien Lê, and Jérôme Pagès. Revealing new relationships among genes by combining external biological knowledge with expression data in an exploratory framework. Florence, Italy, May 7-9 2012. Symposium on Learning and Data Sciences (SLDS).

Julie Josse, **Marie Verbanck**, and François Husson. Pca visualisation improved by regularisation. Oviedo, Spain, December 1-3 2012. ERCIM.

Julie Josse, **Marie Verbanck**, and François Husson. Regularised pca to denoise and visualise data. Bordeaux, France, 8-9 avril 2013. Statlearn.

Marie Verbanck, Julie Josse, and François Husson. Regularised PCA to denoise and visualise data. Toulouse, France, 21-31 mai 2013. Journées de Statistique.

BIBLIOGRAPHIE

O. ALTER, P. O. BROWN et D. BOTSTEIN : Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, août 2000. ISSN 0027-8424, 1091-6490. (pages 44 et 45)

M. ASHBURNER, C. A. BALL, J. A. BLAKE, D. BOTSTEIN, H. BUTLER, J. M. CHERRY, A. P. DAVIS, K. DOLINSKI, S. S. DWIGHT, J. T. EPPIG, M. A. HARRIS, D. P. HILL, L. ISSEL-TARVER, A. KASARSKIS, S. LEWIS, J. C. MATESE, J. E. RICHARDSON, M. RINGWALD, G. M. RUBIN et G. SHERLOCK : Gene ontology : tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1):25–29, mai 2000. (pages 29, 31, 32 et 74)

A. BACCINI, P. BESSE, S. DEJEAN, P. G. P. MARTIN, C. ROBERT-GRANIE et M. SAN CRISTOBAL : Strategies pour l'analyse statistique de donnees transcriptomiques. *Journal de la Société française de statistique*, 146(1-2):5–44, 2005. (page 39)

J. D. BANFIELD et A. E. RAFTERY : Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803, sept. 1993. ISSN 0006341X. (page 29)

T. BARRETT et R. EDGAR : Gene expression omnibus (GEO) : microarray data storage, submission, retrieval, and analysis. *Methods in enzymology*, 411:352–369, 2006. ISSN 0076-6879. PMID : 16939800 PMCID : PMC1619900. (page 34)

J.-P. BENZECRI : *Histoire et préhistoire de l'analyse des données*. Dunod, 1982. (page 42)

T. BLUMENTHAL : Operons in eukaryotes. *Briefings in functional genomics & proteomics*, 3(3):199–211, nov. 2004. ISSN 1473-9550. (page 32)

A. BRAZMA et A. C. CULHANE : Algorithms for gene expression analysis. *In Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley & Sons, Ltd, 2004. ISBN 9780470011539. (page 27)

C. H. BUSOLD, S. WINTER, N. HAUSER, A. BAUER, J. DIPPON, J. D. HOHEISEL et K. FELLENBERG : Integration of GO annotations in correspondence analysis : facilitating the interpretation of microarray data. *Bioinformatics*, 21(10):2424–2429, mai 2005. (page 78)

E. J. CANDÈS, C. A. SING-LONG et J. D. TRZASKO : Unbiased risk estimates for singular value thresholding and spectral estimators. (Submitted), 2012. (pages 132 et 133)

H. CAUSSINUS : *Models and uses of principal component analysis (with discussion)*, p. 149–178. DSWO Press, 1986. (pages 46 et 143)

J. A. CROFT, J. M. BRIDGER, S. BOYLE, P. PERRY, P. TEAGUE et W. A. BICKMORE : Differences in the localization and morphology of chromosomes in the human nucleus. *The Journal of cell biology*, 145(6):1119–1131, juin 1999. (page 33)

J.-P. DANIELS, K. GULL et B. WICKSTEAD : Cell biology of the trypanosome genome. *Microbiology and molecular biology reviews : MMBR*, 74(4):552–569, déc. 2010. ISSN 1098-5557. (page 33)

A. DEAN : In the loop : long range chromatin interactions and gene regulation. *Briefings in Functional Genomics*, 10(1):3–10, jan. 2011. ISSN 2041-2649. (page 32)

C. DÉSERT, M. DUCLOS, P. BLAVY, F. LECERF, F. MOREEWS, C. KLOPP, M. AUBRY, F. HERAULT, P. LE ROY, C. BERRI, M. DOUAIRE, C. DIOT et L. S. : Transcriptome profiling of the feeding-to-fasting transition in chicken liver. *BMC Genomics*, 2008. (page 34)

A. S. DEVONSHIRE, R. ELASWARAPU et C. A. FOY : Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. *BMC genomics*, 11:662, 2010. ISSN 1471-2164. PMID : 21106083. (page 39)

M. B. EISEN, P. T. SPELLMAN, P. O. BROWN et D. BOTSTEIN : Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998. (pages 28, 45 et 73)

B. ESCOFIER et J. PAGÈS : *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*. Dunod, Paris, 2008. ISBN 9782100519323 2100519328. (pages 79 et 101)

Y. ESCOUFIER : Le traitement des variables vectorielles. *Biometrics*, 29(4):751, déc. 1973. ISSN 0006341X. (page 41)

A. FAGAN, A. C. CULHANE et D. G. HIGGINS : A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, 7(13):2162–2171, juin 2007. (page 78)

R. A. FISHER : On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87, jan. 1922. (page 29)

C. FRIGUET, M. KLOAREG et D. CAUSEUR : A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415, déc. 2009. (page 28)

A. GENZ, F. BRETZ, T. MIWA, X. MI, F. LEISCH, F. SCHEIPL et T. HOTHORN : *mvtnorm : Multivariate Normal and t Distributions*, 2013. URL http://CRAN.R-project.org/package=mvtnorm. R package version 0.9-9995. (page 104)

J. A. HARTIGAN et M. A. WONG : Algorithm AS 136 : A k-means clustering algorithm. *Applied Statistics*, 28(1):100, 1979. ISSN 00359254. (page 28)

T. HASTIE, R. TIBSHIRANI, M. B. EISEN, A. ALIZADEH, R. LEVY, L. STAUDT, W. C. CHAN, D. BOTSTEIN et P. BROWN : 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):research0003, août 2000. ISSN 1465-6906. (page 45)

N. S. HOLTER, A. MARITAN, M. CIEPLAK, N. V. FEDOROFF et J. R. BANAVAR : Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences*, 98(4):1693–1698, fév. 2001. ISSN 0027-8424, 1091-6490. (page 45)

N. S. HOLTER, M. MITRA, A. MARITAN, M. CIEPLAK, J. R. BANAVAR et N. V. FEDOROFF : Fundamental patterns underlying gene expression profiles : Simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15):8409–8414, juil. 2000. ISSN 0027-8424, 1091-6490. (page 44)

T. HORINOUCHI, K. TAMAOKA, C. FURUSAWA, N. ONO, S. SUZUKI, T. HIRASAWA, T. YOMO et H. SHIMIZU : Transcriptome analysis of parallel-evolved escherichia coli strains under ethanol stress. *BMC Genomics*, 11(1):579, oct. 2010. ISSN 1471-2164. PMID : 20955615. (page 39)

F. HUSSON, J. JOSSE, S. LE et J. MAZET : *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R*, 2013. URL http://CRAN.R-project.org/package=FactoMineR. R package version 1.25. (page 130)

M. JEANMOUGIN, A. de REYNIES, L. MARISA, C. PACCARD, G. NUEL et M. GUEDJ : Should we abandon the t-test in the analysis of gene expression microarray data : A comparison of variance modeling strategies. *PLoS ONE*, 5(9):e12336, sept. 2010. (page 28)

J. JOSSE et F. HUSSON : Selecting the number of components in pca using cross-validation approximations. *Computational Statistics and Data Analysis*, 56:1869–1879, 2011. (page 144)

J. JOSSE, J. PAGÈS et F. HUSSON : Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, 5(3):231–246, oct. 2011. ISSN 1862-5347, 1862-5355. (page 134)

L. KAUFMAN et P. J. ROUSSEEUW : *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley-Interscience, 1 édn, 1990. ISBN 0471735787. (page 29)

J. C. KWEKEL, V. G. DESAI, C. L. MOLAND, W. S. BRANHAM et J. C. FUSCOE : Age and sex dependent changes in liver gene expression during the life cycle of the rat. *BMC Genomics*, 11(1):675, nov. 2010. ISSN 1471-2164. PMID : 21118493. (page 39)

MATLAB : *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010. (page 130)

L. MIRBAHAI, T. D. WILLIAMS, H. ZHAN, Z. GONG et J. K. CHIPMAN : Comprehensive profiling of zebrafish hepatic proximal promoter CpG island methylation and its modification during chemical carcinogenesis. *BMC genomics*, 12:3, 2011. ISSN 1471-2164. PMID : 21205313. (page 39)

M. OUÉDRAOGO, S. LÊ, M. VERBANCK, C. DIOT et F. LECERF : Identification of gene co-expression structures by multivariate analyses of expression data and comparison with genome interactions. *Nucleic Acid Research (submitted)*, January 2013. (page 107)

J. QUACKENBUSH : Microarray data normalization and transformation. *Nature genetics*, 32 Suppl:496–501, déc. 2002. ISSN 1061-4036. (page 26)

R CORE TEAM : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org`. (page 130)

S. RAYCHAUDHURI, J. M. STUART et R. B. ALTMAN : Principal components analysis to summarize microarray experiments : application to sporulation time series. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 455–466, 2000. ISSN 2335-6936. (pages 44 et 73)

SCILAB ENTERPRISES : *Scilab : Le logiciel open source gratuit de calcul numérique*. Scilab Enterprises, Orsay, France, 2012. URL `http://www.scilab.org`. (page 130)

B. SHARIF et Y. BRESLER : Physiologically improved NCAT phantom (PINCAT) enables in-silico study of the effects of beat-to-beat variability on cardiac MR. *In Proceedings of the Annual Meeting of ISMRM, Berlin*, p. 3418, 2007. (pages 34 et 132)

M. d. TAYRAC, S. LÊ, M. AUBRY, J. MOSSER et F. HUSSON : Simultaneous analysis of distinct omics data sets with integration of biological knowledge : Multiple factor analysis approach. *BMC Genomics*, 10(1):32, jan. 2009. ISSN 1471-2164. (page 78)

C. TER BRAAK : Canonical correspondence analysis : A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167, oct. 1986. ISSN 00129658. (page 80)

M. VERBANCK, J. JOSSE et F. HUSSON : Regularised PCA to denoise and visualise data. *Statistics and Computing (submitted)*, January 2013a. URL http://arxiv.org/abs/1301.4649. (pages 48, 130 et 132)

M. VERBANCK, S. LÊ et J. PAGÈS : A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. *BMC Bioinformatics*, 14(1): 42, fév. 2013b. URL http://www.biomedcentral.com/1471-2105/14/42/abstract. Highly Accessed. (pages 83, 133 et 140)

M. E. WALL, P. A. DYCK et T. S. BRETTIN : SVDMAN—singular value decomposition analysis of microarray data. *Bioinformatics*, 17(6):566–568, jan. 2001. ISSN 1367-4803, 1460-2059. (page 45)

D. WITTEN, R. TIBSHIRANI et T. HASTIE : A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10:515–534, 2009. (page 28)

Y. H. YANG et A. PAQUET : *goTools : Functions for Gene Ontology database*, 2009. R package version 1.34.0. (page 145)

K. Y. YEUNG et W. L. RUZZO : Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, jan. 2001. ISSN 1367-4803, 1460-2059. (page 45)

M. K. S. YEUNG, J. TEGNÉR et J. J. COLLINS : Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168, avr. 2002. (page 45)

B. ZHANG et S. HORVATH : A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4, 2005. (page 73)

R. M. ZIRBEL, U. R. MATHIEU, A. KURZ, T. CREMER et P. LICHTER : Evidence for a nuclear compartment of transcription and splicing located at chromosome domain boundaries. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 1(2):93–106, juil. 1993. (page 33)