Le transcriptome : un domaine d'application pour lesstatistiques, de nouveaux horizons pour la biologie

Anne-Sophie Carpentier

Université d'Evry

Thèse

Présentée pour obtenir le grade de Docteur en Sciences
de l'Université d'Evry

Spécialité : Biologie cellulaire et moléculaire

Par

Anne-Sophie Carpentier

Le transcriptome : un domaine d'application pour les statistiques, de nouveaux horizons pour la biologie

Soutenue le 24 avril 2006, devant le jury composé de :

Claude Millier, Directeur scientifique de l'ENGRREF	Président du jury
Frédérique Hubler Chargé de recherche CNRS à Rennes	Rapporteur
Claude Thermes Chargé de recherche CNRS à Orsay	Rapporteur
Pascale Le Roy Directeur de recherche à l'INRA de Rennes	
Alain Hénaut Professeur à l'Université d'Evry	Directeur de thèse

Résumé

Les mesures des niveaux d'expression de tous les gènes d'un génome requièrent une analyse statistique afin d'obtenir des conclusions fiables. Les biologistes ont du mal à faire un choix dans la foule de méthodes existantes. Afin de déterminer quelle méthode est la plus adéquate pour la problématique abordée, des comparaisons de méthodes d'analyse disponibles sont nécessaires. Actuellement les critères de comparaison se révèlent soit lacunaires ou soit non pertinents du point de vue biologique.

Nous avons introduit un nouveau critère biologique de comparaison des méthodes d'analyse du transcriptome fondé sur une structure des génomes bactériens : les opérons. Les gènes d'un opéron sont généralement transcrits sur un même ARNm. Si un gène d'un opéron bactérien est identifié, les autres gènes de l'opéron devraient l'être également. Nous avons ainsi comparé des méthodes d'analyse appliquées au transcriptome : l'ACP et l'ACI, respectivement analyses en composantes principales et indépendantes, l'ANOVA, analyse de variance, la régression des moindres carrés partiels PLS et différents t-tests. Chaque méthode aborde le nuage de données d'un point de vue différent ce qui donne des résultats complémentaires. Globalement, l'ACI a fourni les meilleurs résultats tant en sensibilité qu'en terme de précision.

Un autre aspect, en plein développement, de l'analyse du transcriptome est la méta-analyse de données d'origines diverses malgré les biais inhérents à cette technologie. Généralement ces méta-analyses visent à préciser les résultats concernant des gènes différentiellement exprimés ou co-exprimés. Elles ouvrent également la possibilité d'étudier de nouveaux champs en biologie. Nous avons utilisé des données de transcriptome indépendantes afin d'étudier l'organisation de l'expression des gènes et, ainsi, celle du chromosome bactérien. L'étude du transcriptome de trois bactéries, *B. subtilis, E. coli et S. meliloti* a révélé des corrélations d'expression à longue distance valables quel que soit le gène étudié. Les structures en opéron se manifestent clairement au travers de cette étude, qui a également permis de préciser que la co-expression de gènes proches s'étend au-delà des opérons dans une région qui se répand jusqu'à une centaine de gènes.

Pour conclure, l’analyse du transcriptome n’a pas réellement nécessité la mise au point de méthodes d’analyse statistiques spécifiques. Cependant, elle permet d’aborder de nouveaux horizons dans la biologie, et notamment l’organisation chromosomique du génome bactérien.

Remerciements

Je tiens à remercier Alain Hénaut qui m'a guidée tout en me laissant une liberté d'action sans égale. A chaque fois que tu me proposais de faire certaines analyses complémentaires pour lesquelles je doutais des résultats, tu as toujours eu raison. Les résultats ne seraient pas là sans toi. Je tiens aussi à te remercier pour tout ce que j'ai découvert pendant cette thèse qui a eu lieu grâce à toi.

Je tiens à remercier Alex Grossmann qui avait toujours des idées brillantes et sans qui les résultats ne seraient sans doute pas aussi beaux. Tu as partagé avec moi ta passion pour la recherche et tes élans enthousiastes lorsque les résultats étaient là. J'admire grandement ton ouverture d'esprit et ta curiosité exceptionnelles. Je remercie Claudine Landès-Devauchelle qui m'a poussée vers la programmation et m'a donné le courage qu'il me manquait parfois, qui a su m'écouter et me pousser vers le C. Je remercie Yolande Diaz pour m'avoir soutenue lors de ma thèse. Ta présence a toujours été un réel plaisir et un partage d'expériences parfois tristes et de bons fous rires. Ton support m'a été très important. Je remercie Jean-Loup Risler pour m'avoir conseillée dans la présentation des résultats et surtout dans la rédaction des articles. J'espère m'être un peu inspirée de ta capacité à rédiger clairement. Je remercie Gucki Hénaut pour ton aide dans la rédaction mais surtout pour ta bonne humeur, ta simplicité et ta générosité. J'espère pouvoir m'inspirer de ta façon d'aborder la vie. Merci à Gilles Didier pour m'avoir écrit des programmes bien utiles et à Frédérique Hubler pour m'avoir permis d'utiliser ses données de transcriptome. Je remercie également Bruno Torrésani pour les discussions que nous avons eu sur le modèle de l'organisation chromosomique bactérienne. Je remercie Alexandra Louis qui a toujours été là quand j'avais des problèmes informatiques et qui me procurait régulièrement des informations sur le sujet. Je remercie Virginie pour les clés que j'oubliais souvent ou dont je ne disposais pas. Je te remercie également pour m'avoir écoutée et soutenue. Merci à Carène pour tes conseils avisés et son soutien lors de la rédaction. Je remercie également Pierre, Sophie, Marie-Odile et Yvan qui étaient là pendant ma thèse et sans qui le laboratoire n'aurait pas été le même.

Je remercie Bernard Prum de m'avoir conseillée et soutenue lors de ma thèse et de m'avoir recueillie dans son laboratoire et Claude Millier pour m'avoir permis de faire cette thèse.

Je remercie tous les membres de mon jury d'avoir accepté de participer à ma soutenance de thèse ainsi que l'ENGREF et le ministère de l'agriculture et de la pêche qui m'ont financée.

Je remercie mes amies qui m'ont soutenue en permanence et qui m'ont apporté leur passion du travail et de la vie : Séverine et Emmanuelle. Je remercie également Stéphanie, Céline, Pauline, Aurélie, Sarah d'avoir partagé de très bons moments. Je remercie mon frère Laurent qui s'est inquiété pour moi et qui a partagé mes problèmes informatiques, mon frère David qui sans aucun doute a été à mes débuts un exemple pour moi et à travers qui j'ai découvert le monde de la recherche.

Je remercie enfin et surtout mes parents qui m'ont apporté leur soutien indéfectible, leur aide et surtout leur amour. Ils m'ont donné un vrai sens des valeurs et du travail que j'espère garder et une vision sereine de la vie que j'aimerai toujours partager.

Enfin je dédie cette thèse à Emeric Dietze qui me fait le plaisir de partager ma vie et qui a subi les mauvais moments avec une sérénité sans égale. Il m'a corrigé mes fautes d'orthographe mais surtout soutenue et supportée en permanence et m'a poussé tout au long des difficultés rencontrées. J'ai un peu appris à écrire et je suis plus forte grâce à toi.

Introduction

L'étude du transcriptome est une partie fondamentale de « l'ère de la post-génomique », qui fait suite à « l'ère du séquençage » et à « l'ère de la génomique ». Elle a pris naissance à la fin des années 1990. Elle permet l'étude dynamique des gènes d'organismes modèles identifiés par la génomique. En 1996, lors de l'obtention du génome complet de la levure, 6200 phases ouvertes de lecture sont identifiées. Alors que cet organisme était fortement étudié auparavant, seulement un quart de ces gènes pouvait être associé à une fonction connue ou putative à partir d'homologies de séquence. Ainsi, malgré la connaissance de la séquence d'un organisme, une grande partie du monde génomique reste encore inexplorée [1, 2]. Depuis la fin des années 90, une nouvelle ère de la biologie porte sur une vision dynamique globale du fonctionnement cellulaire : la post-génomique.

Cette thèse porte sur la génomique fonctionnelle et plus particulièrement sur l'étude des niveaux d'expression des gènes. Ce niveau d'étude est intermédiaire entre la génomique et la protéomique et représente le premier niveau d'intégration entre les facteurs environnementaux et le génome (cf. Figure 1). Les ARN messagers synthétisés fournissent une base pour l'adaptation de la cellule à son environnement *via* la synthèse des protéines. Ils pilotent ainsi la réalisation de macro-phénotypes complexes comme la morphologie et le comportement.

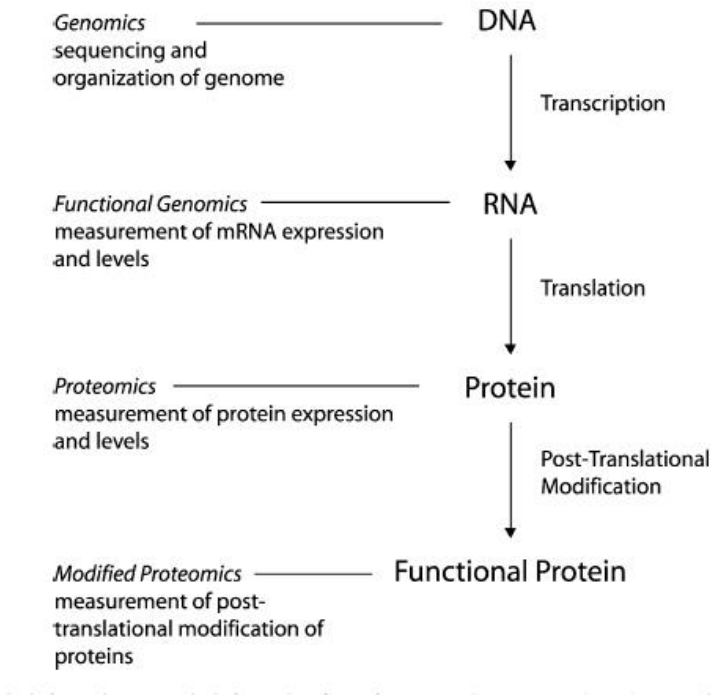


Figure 1 : Différents niveaux d'étude du fonctionnement cellulaire (tirée de [3])

Récemment, des techniques de mesure des niveaux d'expression de l'ensemble des gènes d'un organisme se sont développées, notamment la technique très utilisée des puces à ADN. Les niveaux d'expression relatifs de l'ensemble du génome sont évalués pour chaque condition expérimentale. L'approche diffère avec les approches biologiques classiques puisque l'on passe d'une approche fondée sur une hypothèse à tester à une approche indépendante de modèle, c'est-à-dire relativement exploratoire [4].

Le transcriptome requière des analyses statistiques qui dépendent généralement des questions posées. Or, la communauté de biologistes n'est généralement pas formée aux statistiques. Un des défis majeurs de la communauté biologique et bioinformatique est donc d'adopter une vision plus statistique et d'interagir avec des statisticiens [4]. Le biologiste de laboratoire et le théoricien ont besoin de faire des efforts concertés afin de concevoir des expériences qui peuvent à la fois être réalisées et analysées. Lors de ce doctorat, nous avons été confrontés aux difficultés éprouvées par les biologistes et, parmi celles-ci, au choix de la méthode d'analyse et l'étude massive de données.

La technique des puces à ADN est caractérisée par des difficultés de reproductibilité dues à de nombreux biais techniques ainsi que par les jeux de données obtenus qui comportent de nombreuses variables à étudier (les gènes) pour un petit nombre de mesures. L'analyse de ces données a fait l'objet de la publication d'un très grand nombre d'articles qui tendent à prendre en compte les spécificités des données ainsi que les buts des analyses pratiquées. Au final, plusieurs centaines de méthodologies différentes ont été décrites et aucune n'a obtenu le consensus général. Le choix de la méthode d'analyse adéquate s'avère d'autant plus difficile que les comparaisons de ces méthodes reposent, entre autre, sur des modélisations de données ou des connaissances *a priori* incomplètes. En clair, il n'existe pas de jeu de données test dans lequel les gènes à détecter comme différentiellement exprimés ou corrélés sont connus. Enfin, les conclusions de ces comparaisons sont souvent contradictoires et dépendent du critère et du jeu de données employés.

Parallèlement, le nombre grandissant de données de transcriptome librement accessibles ouvre la voie à la méta-analyse des puces. Comme cette technique dépend fortement de la technique employée mais surtout des facteurs biologiques expérimentaux, les résultats obtenus à partir d'un seul jeu de données sont difficilement généralisables à un phénomène global comme, par exemple, la compréhension du cancer. Depuis 2002, différentes études ont porté sur l'analyse de plusieurs jeux de données obtenus à partir de différentes plates-formes afin d'identifier des résultats reproductibles et sans doute généralisables (gènes impliqués dans un processus, groupes de gènes aux profils d'expression cohérents). Par ailleurs, la méta-analyse peut permettre l'accès à des résultats plus globaux comme l'analyse de régularités d'expression au sein du génome indépendamment des conditions expérimentales. Pour les bactéries, l'organisation chromosomique de l'expression n'avait pas fait l'objet de méta-analyse du transcriptome. Seule une étude du transcriptome de *Escherichia coli* a été effectuée dans ce but [5] . L'interprétation des résultats a été réalisée en fonction des facteurs expérimentaux (dans ce cas deux conditions expérimentales) : l'implication de protéines (les gyrases) dans le repliement du chromosome bactérien.

Cette thèse a porté sur l'étude et l'utilisation de méthodes d'analyse classiques afin de tirer le maximum d'informations des données de puces de transcriptome. Nous avons abordé l'analyse des puces sur deux plans :

- la détermination d'un critère de comparaison de méthodes statistiques indépendant de la méthode utilisée, avec une pertinence biologique également indépendante des conditions expérimentales. Nous avons utilisé les opérons bactériens dont les phases ouvertes de lecture (ORFs) sont généralement présents sur un même ARN messager.
- la poursuite de notre analyse du transcriptome bactérien vers l'organisation de l'expression des gènes. Pour cela, une méta-analyse sur les deux bactéries modèles *Escherichia coli* et *Bacillus subtilis* a été effectuée sur des jeux de données indépendants récupérés sur internet. Nous en avons déduit un modèle d'organisation chromosomique qui a été confronté aux données d'une autre bactérie *Sinorhizobium meliloti,* qui possède plusieurs éléments génomiques circulaires.

La première partie de ce rapport portera sur une bibliographie et une introduction à la technique des puces à ADN en soulignant différents biais existants. Son but est à la fois de comprendre les données de transcriptome mais également de percevoir les points sensibles de la technique. Elle est complétée par un article de revue bibliographique qui détaille les différents types d'analyse des puces.

La deuxième partie porte sur l'étude que nous avons menée afin de déterminer un critère de comparaison des nombreuses méthodes d'analyse utilisées en transcriptomique. Nous abordons le critère fondé sur les opérons avec l'article correspondant complété par les résultats d'une communication orale à la conférence internationale sur les modèles stochastiques appliqués et l'analyse des données (ASMDA) de Brest en 2005.

La dernière partie correspond à l'analyse massive de données d'expression et, plus particulièrement, nos résultats sur les régularités de corrélations d'expression des bactéries que nous appelerons la structure d'expression bactérienne. Nous avons observé des régularités d'expression à grande échelle, valables quel que soit le gène considéré. En clair, pour n'importe quel gène, son expression sera sans doute corrélée avec celles de gènes situés à une distance précise de lui sur le chromosome. Inversement, si un gène est exprimé, les gènes situés à une autre distance qui dépend de la longueur du chromosome, seront réprimés. Cette partie présente un article publié sur l'étude de *E. coli* et de *B. subtilis* ainsi que des résultats d'un article en préparation sur *S. meliloti*.

Note : Dans l'ensemble de ce manuscrit, le terme puce à ADN correspondra à l'ensemble des techniques permettant l'acquisition des données de transcriptome à part la méthode SAGE. Le terme français devrait être « micro-réseaux », mais le terme le plus couramment utilisé reste le terme « puce ». Ce manuscrit n'abordera pas les puces protéiques. Pour approfondir ce point, on peut se référer à la revue bibliographique [6]. Cette technique n'est pour l'instant pas couramment utilisée car elle présente de nombreux problèmes supplémentaires du point de vue technique.

Table des matières

1 Etat des lieux des puces à ADN et de leurs méthodes d'analyse

Pour comprendre l'explosion des méthodes statistiques d'analyse des puces et les étudier, il faut savoir :

- Quel a été l'apport des puces dans la compréhension des mécanismes biologiques.
- A quoi correspondent biologiquement et techniquement les puces.
- Comment les données sont traitées et analysées.

1.1 Apport des puces à ADN pour mesurer le niveau d'expression des gènes

Autres techniques de mesure des niveaux d'expression

Avant les puces, il existait déjà des techniques pour mesurer les niveaux d'expression différentiels des gènes. Certaines de ces techniques sont toujours utilisées parallèlement à l'utilisation des puces.

- Les Northern Blot [7]: une technique qui permet de détecter la présence d'ARN messagers (ARNm) spécifiques mais également des ARN non codants comme les petits ARN et les ARN ribosomaux à l'aide de sondes marquées. Les ARN messagers d'un échantillon sont séparés par électrophorèse. La mise en présence du résultat de l'électrophorèse avec une sonde radioactive d'ADN complémentaire (ADNc) de l'ARNm recherché entraîne la détection ou non d'un ARN. La présence de l'ARN est révélée par autoradiographie. Cette technique permet de mesurer l'expression relative d'au plus 20 gènes à la fois.

- L'analyse d'expression différentielle [8]: une technique de détection des gènes qui sont exprimés uniquement sous certaines conditions. Elle consiste à isoler et à comparer les ARNm amplifiés par PCR (*polymerase chain reaction*) provenant d'au moins deux populations cellulaires. Le criblage se fait *via* des gels d'électrophorèse et se fonde donc sur les différences de longueurs d'ARNm. Les ARNm sont ensuite identifiés grâce au séquençage de l'ARNm prélevé. Cette technique présente des inconvénients comme de nombreux faux positifs générés par la PCR

(environ 50%) et un biais en faveur des ARN abondants. L'expression différentielle obtenue doit donc être validée par une autre technique.

- SAGE ou *Serial Analysis of Gene Expression* [9] : cette technique correspond à une version accélérée du séquençage des EST. Elle est fondée sur le fait que de petites séquences seraient suffisantes pour identifier un gène transcrit du moment que l'on connaît la position de ces séquences dans le transcrit. La technique comprend une extraction des ARN et une transcription reverse en ADN double brins. Une enzyme de restriction coupe ensuite les ADN obtenus en fragments de dix à quinze paires de bases (pb). Ceux-ci sont ligaturés ensemble et amplifiés par PCR puis séquencés. Comme chaque petit morceau de quinze pb correspond théoriquement à un ARNm transcrit, l'identité des transcrits ainsi que leur abondance sont alors connus [10]. Elle présente l'avantage que l'ensemble des gènes exprimés sont quantifiés si un nombre suffisant de fragments ont été séquencés. Cependant, l'identification des gènes dépend des séquences déjà disponibles dans les bases de données. Le choix de l'enzyme de restriction peut influer sur les résultats et restreindre la présence de certains transcrits. Van Hal *et al.* [11] caractérisent cette technique comme laborieuse, qui nécessite des procédures de préparation des échantillons complexes, et qui demande un séquençage d'ADN considérable.

- Le dot blot [12] dont le principe est à l'origine des puces à ADN, demande une quantité de matériel relativement considérable à cause de la taille des filtres.

Moody [13] décrit en détail la plupart de ces méthodes et recense leurs avantages et leurs inconvénients.

Actuellement, une technique est couramment utilisée afin de mesurer le niveau d'expression d'un gène : la RT-PCR quantitative (*Reverse Transcribed Polymerase Chain Reaction*). Cette technique est très sensible et permet la détection des ARNm présents en un seul exemplaire. Elle n'est cependant pas utilisée pour mesurer l'expression de l'ensemble des gènes d'un organisme simultanément.

L'ensemble des techniques décrites ci-dessus permet soit l'analyse relativement précise de quelques gènes à la fois soit la mesure plutôt qualitative d'un grand nombre de gènes. Seule la méthode SAGE présente la possibilité d'obtenir les niveaux d'expression de l'ensemble des ARNm de manière quantitative. Cependant, elle requiert un travail très important de séquençage et est actuellement largement devancée par les puces à ADN.

La conception des puces à ADN

Les puces à ADN correspondent à un ensemble de techniques permettant la mesure relative simultanée des niveaux d'expression de milliers de gènes, voire de l'ensemble du génome d'un organisme.

Comme la plupart des techniques précédemment citées, elles se fondent sur le principe d'hybridation : deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible sous l'action de la chaleur et de la concentration saline du milieu. Les conditions expérimentales définissent le degré de complémentarité des deux séquences nécessaire à l'hybridation. Les puces sont une miniaturisation du système classique de reverse dot blot [12].

Même si l'utilisation courante des puces à ADN date des années 2000, l'introduction du principe de cette technique est beaucoup plus ancienne. A la fin des années 1980, les immunodosages ou tests ELISA étaient couramment utilisés pour doser la présence de certaines molécules dans des échantillons. Ce test reposait sur des anticorps spécifiques marqués généralement par radioactivité. Ekins *et al.* [14] ont proposé d'utiliser des marquages fluorescents à la place du marquage radioactif et de réaliser ces réactions immunologiques grâce à des spots disposés sur un support solide. Cela ouvrait la possibilité de construire des immunodosages de dizaines ou centaines de molécules à la fois avec l'aide de techniques de balayage *via* les lasers. Le fait que ces micropuces présentaient la même sensibilité que les tests plus encombrants était une réelle nouveauté. Par la suite, ils proposèrent également l'utilisation de deux marqueurs [15, 16] afin de doser les molécules présentes ainsi que l'utilisation du ratio entre les intensités obtenues par ces deux marquages.

Toutes ces avancées ont été reprises afin de mettre en place les puces à ADN. La différence entre les « puces à immunodosage » décrites par Ekins et collaborateurs et les puces à ADN reposent sur les molécules analysées. Dans le premier cas, il s'agit d'antigènes dosés *via* des anticorps correspondants, dans l'autre il s'agit d'ARNm dosés *via* leurs séquences complémentaires fixées sur le support.

Le premier article au sujet de puces pour l'étude de l'expression des gènes a été publié en 1995 par Schena *et al.* [17] en parallèle de la publication de la méthode SAGE [9]. Schena et collaborateurs ont développé une puce de 45 gènes d'*Arabidopsis thaliana*. La double fluorescence était utilisée afin de déterminer les différences de niveaux d'expression entre différents organes de la plante. Lockhart *et al.* [18] ont présenté en 1996, la technique de puces à oligonucléotides permettant de détecter plusieurs dizaines de milliers de gènes à la fois.

Cette technique a soulevé de nombreux espoirs quant à la compréhension globale de l'expression des gènes : conditions d'expression, régulation, fonction, … Cet engouement a conduit à une croissance exponentielle du nombre de publications référencées par Pubmed qui contiennent les noms «microarray», « DNA chip », « expression array », « gene chip » ou « gene array » (Figure 2). Les puces sont utilisées pour de nombreux types d'applications dont la principale consiste à mesurer les niveaux d'expression.

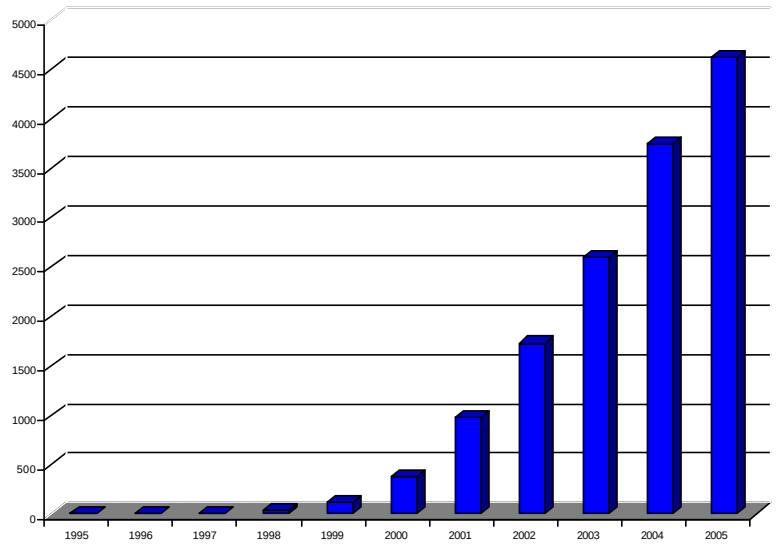


Figure 2 : Croissance exponentielle du nombre de publications concernant les puces à ADN

Applications des puces pour la génomique

De nombreuses applications des puces se développent pour l'étude du génome des organismes en dehors de la simple expression des gènes. La description ci-dessous reste succincte et non exhaustive puisque le reste de cette thèse ne porte pas sur ses applications. On pourra se référer à la revue de Mockler *et al.* [19] pour de plus amples détails.

Etude du polymorphisme

Au départ, les puces n'ont pas été mises au point afin d'étudier le niveau d'expression des gènes mais afin de séquencer les génomes. Grâce aux puces à ADN, différents aspects génomiques sont étudiés. Ces expériences requièrent parfois l'utilisation de puces génomiques, c'est-à-dire des puces où sont représentées à la fois des séquences de gènes et des séquences intergéniques.

En 1988-89 Fodor *et al.* [20] et Southern *et al.* [21] de la société Affymetrix ont développé une méthode de séquençage par hybridation (*sequencing by hybridization* SBH). Le gène étudié est considéré comme un ensemble de plusieurs séquences chevauchantes dont la détermination simultanée puis l'assemblage permettent de reconstituer la séquence [22] (cf. Figure *3*). Pease *et al.* en 1994 [23] ont complété cette technologie. Afin d'avoir des détails sur les techniques existantes et sur leurs utilisations, on peut se référer à la revue [22]. Les puces à ADN sont actuellement utilisées pour le re-séquençage afin d'élucider les différences entre des séquences d'origines différentes. Plus particulièrement, on recherche des mutations entre différents individus ou populations comme le séquençage de différentes souches de streptocoques afin d'adapter les traitements aux infections [24].

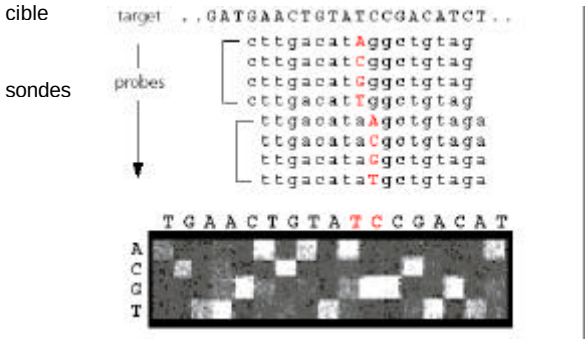


Figure 3 : Principe du séquençage par hybridation (image tirée de [25])

Des puces à oligonucléotides ont été fabriquées afin d'identifier les polymorphismes sur une base spécifique (*single nucleotide polymorphism* ou SNP). Les oligonucléotides sont organisés en tétrades au sein desquelles une séquence est strictement identique à celle de type sauvage alors que les trois autres sont caractérisées par une substitution de base localisée au milieu de la séquence. Ces puces [25] permettent la détection de différents allèles ainsi que leurs positions (génotypage) mais sont inadaptées à l'étude de polymorphisme long comme des délétions ou des insertions. Plusieurs génotypages ou criblages de mutations ont été effectués grâce à cette technique [26-28]. La revue [29] récapitule l'ensemble des techniques utilisées actuellement pour la détection de SNP dont les puces à ADN.

Les puces peuvent également être utilisées afin d'étudier la séquence d'organismes proches de celui pour lequel elles ont été conçues. Jusqu'à présent, la séquence complète du génome n'est disponible que pour un petit nombre des systèmes modèles [30]. Pour les organismes non modèles, l'approche la plus accessible est l'utilisation de puces à ADNc afin d'établir des différences de séquences (insertion/délétion) avec des organismes proches. Cette application est également étendue à la recherche de régions délétées ou insérées entre différentes souches ou populations. Lashkari *et al.* [2] ont présenté des puces permettant la comparaison génomique de deux souches de levure. Après hybridation de l'ADN génomique marqué, les gènes pour lequel le signal est extrêmement faible sont d'éventuels gènes délétés ou extrêmement divergents entre différentes souches. Une autre technique, les puces CGH (*comparative genome hybridization*), puces génomiques, permettent également l'identification de grandes délétions/ insertions. Comme des cancers peuvent se caractériser par des délétions/ insertions ou amplifications de certaines séquences, les puces CGH ont été utilisées sur

des lignées cellulaires tumorales [31, 32] et ont permis de détecter des séquences amplifiées et délétées dans différents types de cancer.

Analyse des mécanismes de régulation d'expression (hybridation d'ADN)

Toujours grâce à l'hybridation de l'ADN aux puces, des études se sont focalisées sur une échelle plus réduite, le gène et ses environs. Les puces ont ouvert la voie à des cartographies des motifs de régulations de l'ADN à grande échelle. On cherche à identifier des régions d'ADN se liant avec les facteurs de transcription. Pour cela, la chromatine liée à des protéines est immunoprécipitée grâce à des anticorps spécifiques de la protéine étudiée. La séquence est ensuite identifiée grâce à des puces génomiques. Cette technique, ChIP-on-chip, et les découvertes réalisées grâce à elle sont décrites en détail dans les revues [33, 34]. L'étude de la liaison de deux facteurs de transcription [35] a montré leur implication dans l'activation de gènes qui font partie de deux voies métaboliques différentes: la synthèse de la membrane cellulaire et la réplication et la réparation de l'ADN. Cette spécialisation des facteurs de transcription permet d'expliquer la régulation de processus cellulaires indépendants. Globalement, les puces génomiques permettent donc d'aborder le système de régulation d'expression de manière complètement nouvelle. Ainsi, une grande partie des sites de fixation de facteurs de régulation identifiés grâce à cette technique se situent en dehors des régions promotrices prédites jusqu'alors [19, 36].

Une autre application concerne la régulation de l'expression des gènes *via* la méthylation des cytosines de l'ADN. Plusieurs techniques sont utilisées afin d'identifier les sites de méthylation de l'ADN (cf. la revue [19]). Globalement, une réaction préalable permet de différencier les cytosines où l'ADN est méthylé de celles où il ne l'est pas. Des enzymes de clivage pour les ADN non méthylés ou la conversion des cytosines en uracile grâce au bisulfate de soude sont deux types de réactions utilisées. Ces morceaux d'ADN méthylés sont ensuite identifiés grâce à l'hybridation sur des puces.

Analyse au niveau de l'ARN et des mécanismes d'expression

Les puces à ADN génomique permettent de découvrir de nouveaux gènes [19] et de définir leur structure (intron/exons pour les eucaryotes). Des puces génomiques qui représentent soit l'ensemble du génome soit des portions du génome régulièrement espacées permettent la détection des portions d'ADN transcrites. Pour l'homme et le génome des plantes, les puces génomiques révèlent que certaines régions considérées comme non codantes sont effectivement exprimées [37-39]. Parmi les séquences exprimées, environ 50% étaient annotées comme étant des gènes alors que 50% ne présentaient aucune annotation codante. La revue de Johnson *et al.* [40] permet d'approfondir cette question et montre notamment qu'une partie des séquences transcrites sont en fait des ARN non traduits ou correspondent à des transcrits anti-sens.

Une autre application des puces à ADN est l'analyse des transcrits alternatifs difficilement prédits par les programmes actuels de la bioinformatique. Cela a été appliqué chez la levure [41]. La technique n'est pas encore sans défaut. On peut se référer à la revue afin de connaître ses limites et les différentes applications existantes [42].

L'étude des séquences cibles des protéines liant l'ARN permet de comprendre les mécanismes de modifications post-transcriptionnelles grâce aux puces. Cette analyse peut être couplée avec l'épissage alternatif afin de comprendre comment les différents ARNm sont synthétisés [19]. Le principe est simple et ressemble à celui de ChIP-on-chip : les séquences d'ARN qui se lient avec une protéine sont purifiées par immunoprécipitation ou par colonne d'affinité. La puce à ADN permet ensuite d'identifier ces séquences. La revue [43] référence l'ensemble des applications des puces quant à l'identification des séquences d'association entre l'ARN et des protéines.

Enfin, de nouvelles puces sont actuellement développées afin d'étudier l'effet des *ARN interference* (ARNi) sur les cellules [44]. Des lots d'ARNi sont fixés sur la puce puis mis en présence de cellules à transfecter. Le but est d'observer les phénotypes obtenus sur les cellules transfectées par un ARNi particulier.

Applications des puces pour l'étude de l'expression des gènes

Actuellement, la principale utilisation des puces à ADN est de mesurer les niveaux d'expression relatifs des gènes dans différentes conditions. Si la génomique permet de connaître la structure du génome et les différents gènes existants, elle ne permet d'aborder la fonction des gènes que *via* les similitudes de séquences, c'est-à-dire de manière statique et dépendante des connaissances déjà acquises. Or, comme le précise Lipshutz *et al.* en 1999 [25], pour comprendre la fonction d'un gène il est utile de savoir quand et où il est exprimé et dans quelles circonstances son niveau d'expression est affecté. Les puces permettent d'aborder ces problèmes. Néanmoins outre les études concernant un gène précis, elles permettent d'aborder l'étude des réseaux fonctionnels existants.

Au départ, les puces engendraient de nombreux espoirs. En 1999, Brown et Botstein [1] parlent de nouvelles cartes génomiques fondées sur les données d'expression, qui permettraient de comprendre les fonctions des gènes ou leur régulation. Pour eux, dans un avenir proche, c'est à dire à peu près de nos jours, les effets des mutations pour chaque gène de la levure seraient connus. Tel n'est pas le cas. Si les puces constituent un outil qui fait avancer la compréhension du fonctionnement des gènes, il ne répond pas entièrement aux attentes formulées par Brown et Botstein du fait de ses limites.

Malgré cela, de nombreux aspects de l'expression des gènes ont été abordés grâce aux puces comme nous allons le détailler ci-après. On peut classer les applications existantes des puces en différents types [45] : la comparaison des niveaux d'expression de gènes selon différentes conditions, l'identification de gènes fonctionnellement liés, la recherche de gènes discriminants grâce au clustering, l'approche des réseaux d'interaction géniques grâce aux séries temporelles.

Comparaison des niveaux d'expression des gènes selon différentes conditions

La première utilisation des puces pour les niveaux d'expression [17] a été l'analyse des différences d'expression de 45 gènes entre les feuilles et les racines d'*Arabidopsis thaliana*. Depuis, de nombreuses expériences de recherche de gènes différentiellement exprimés entre plusieurs

conditions ont été réalisées. Le but est de comprendre certains phénomènes biologiques et d'identifier les gènes impliqués dans ces processus.

Plusieurs phénomènes ont été étudiés : la différenciation et la maturation de cellules dendritiques humaines et les gènes impliqués dans ces phénomènes [46]. Clark *et al.* [47] ont identifié des gènes dont l'expression induit le passage d'une cellule tumorale en métastase chez la souris. DeRisi *et al.* [48] se sont intéressés aux différences d'expression entre des lignées de cellules cancéreuses humaines et des lignées mutantes qui suppriment les tumeurs. Hedenfalk *et al.* [49] ont étudié les différences d'expression entre des mutations de BRCA1 et BRCA2 dans des cas de cancer du sein.

En plus de ces études du cycle cellulaire et du cancer, certaines applications visent à analyser le comportement des gènes face à des modifications environnementales notamment le stress.

Les puces ont également permis d'identifier les gènes dont l'expression est gouvernée par différents régulateurs, par les brassinostéroïdes et les gibbérellines pour le riz [50] ou encore d'identifier les mécanismes de régulation des défenses d'*Arabidopsis thaliana* [51].

Différents métabolismes ont encore été étudiés comme le mécanisme de l'assimilation du soufre chez *Bacillus subtilis* [52].

Identification de gènes fonctionnellement liés

Une des applications les plus courantes des puces est l'étude de gènes dont les niveaux d'expression sont corrélés. Eisen *et al.* [53] ont mis au point une méthode permettant d'identifier les gènes partageant le même profil d'expression. L'hypothèse suppose que des gènes co-exprimés sont impliqués dans un même processus cellulaire ou voie métabolique. Il serait ainsi possible d'identifier les fonctions de gènes jusqu'alors inconnues en analysant les fonctions des gènes ayant le même profil d'expression. La revue de Niehrs and Pollet [54] recense, chez les eucaryotes, différents groupes de gènes qui ont été identifiés comme co-exprimés en 1999 et qui partagent une fonction commune.

Une autre hypothèse est que les gènes identifiés comme co-exprimés seraient régulés par des facteurs de régulations communs et auraient donc, dans leur promoteur, des sites de fixation identiques.

Plusieurs études combinent ces deux hypothèses comme celle de Chu *et al.* [55] en 1998 sur les différentes étapes de la sporulation chez la levure. Les différents prélèvements ont permis de distinguer différentes phases de la sporulation et de caractériser les gènes associés. Lors de leur travail, ils ont tenté d'identifier des séquences nucléiques communes aux gènes identifiés afin de découvrir des sites de régulations communs. Des études similaires ont été réalisées afin d'étudier le cycle cellulaire de la levure [56, 57], le cycle circadien d'*Arabidopsis thaliana* [58]. Les cycles cellulaires de fibroblastes humains [59] ou de cellules cancéreuses humaines [60] ont également été étudiés afin d'identifier de nouveaux gènes impliqués dans différentes phases du cycle.

En plus des séries temporelles, d'autres études portent sur la compréhension de la réponse d'organismes à différentes conditions comme l'étude des réponses de la levure à différents stress environnementaux [61].

Recherche de gènes discriminants

Les puces peuvent également être un outil de diagnostic afin de discriminer entre différents types cellulaires. Dans le domaine clinique notamment, la recherche vise à identifier des gènes marqueurs qui permettent de réaliser un diagnostic fiable de maladies comme le cancer. Alizadeh *et al.* [62] en 2000 ont démontré qu'il existait des signatures d'expression différentes pour des cellules de lymphomes qui proviennent soit de patients qui répondent à un traitement soit de patients qui ne répondent pas et succombent. Van't Veer *et al.* [63] ont ainsi utilisé les puces afin de déterminer un ensemble de gènes dont les profils d'expression « prédisent » les possibilités de métastases pour des tumeurs du sein. Plusieurs études [64, 65] ont identifié des groupes de gènes pouvant prédire la possibilité de survie pour des patients atteints de cancer des poumons.

L’étude de l’expression des gènes permet également de distinguer des types cellulaires ou de cancer difficilement identifiables par les critères physiologiques classiques [66-68]. Les puces permettent donc de faire des diagnostics de maladie et d’adapter les traitements aux différents types de maladies.

Approche des réseaux d’interaction géniques

Les puces à ADN mettent à disposition des jeux relativement larges de données d’expression d’une grande partie des gènes d’un organisme. Parallèlement à l’étude de groupes de gènes co-exprimés se sont développées des méthodes qui permettent d’inférer les réseaux d’interaction géniques (régulation) d’un organisme. Ces études mènent à une visualisation du réseau plus ou moins dense. Rung *et al.* [69] ont analysé un important jeu de données chez la levure.

1.2 La partie technique et biologique

Le principe des puces

Le principe de la puce à ADN est l'hybridation spécifique de deux molécules présentant la même séquence. Sur une surface de quelques centimètres carrés, des fragments synthétiques d'ADN (les sondes) sont greffés et espacés de quelques micromètres. Les sondes sont regroupées en spots représentatifs de chacun des gènes étudiés. Ce micro-dispositif est ensuite mis au contact des acides nucléiques à analyser, au cours de l'étape d'hybridation. Ces acides nucléiques, appelés cibles, correspondent aux ARNm ou aux ADNc préalablement couplés à un marqueur fluorescent ou radioactif. Ce contact entre cibles et sondes conduit à la formation d'hybrides qualifiés par leurs coordonnées et quantifiés grâce à la lecture des signaux radioactifs ou fluorescents. Les sondes sont toujours en excès par rapport aux cibles.

L'utilisation des puces à ADN se décompose en différentes étapes qui suscitent plus ou moins l'attention des biologistes : le choix d'un plan expérimental, le choix ou la fabrication de la puce utilisée, l'extraction des ARNm, la synthèse et le marquage des ADNc, l'hybridation, la lecture des données, les étapes de normalisation et d'analyse et enfin la confrontation avec des données externes. Dans cette partie, nous allons nous attacher à la description des différentes étapes de manipulation et d'analyse des résultats ainsi qu'à celle des biais éventuels qui en résultent. Dans le chapitre suivant, nous nous concentrerons sur les particularités des données et leur analyse.

Nous rappelons que dans l'ensemble de ce manuscrit le terme de puces à ADN correspondra à l'ensemble des techniques permettant l'acquisition des données de transcriptome et comprendra donc les micro-réseaux, les filtres à membranes et les puces à oligonucléotides. Toutefois, dans la description des différents supports existants, nous ferons la distinction entre les différentes techniques.

Le choix du plan d'expérience

Le plan d'expérience est une des étapes les plus importantes [70-75] dans la mise en place de la mesure du transcriptome et ce, d'autant plus que les puces génèrent de grandes quantités de données qu'il faut ensuite analyser. Malheureusement, cette étape est fréquemment négligée [73] lors des expériences et se résume souvent au choix d'un protocole expérimental.

Le choix du plan expérimental dépend des questions posées lors de l'expérience et des hypothèses qui en résultent [73, 76]. Il comprend :

- la détermination des conditions expérimentales étudiées et donc des facteurs pris en compte dans l'étude
- leur agencement.

Première étape : définir précisément le but de l'expérience et le(s) facteur(s) d'intérêt

La première question à se poser lors de la mise en place d'une expérience de transcriptome est de définir précisément son but, ce que l'on recherche. Si cette interrogation semble simple en théorie, elle n'est pas si évidente en pratique. Définir précisément quel est le facteur d'intérêt n'est pas une chose aisée. Parfois la question posée correspond à une combinaison de facteurs.

Par ailleurs, si le but est de comprendre un mécanisme physiologique particulier, le choix des conditions étudiées influencera grandement les résultats. Ainsi, si des conditions relativement éloignées sont choisies (exemple un type de cancer *versus* en bonne santé), les niveaux d'expression de nombreux gènes seront affectés : les gènes directement impliqués dans le mécanisme à étudier (gènes impliqués dans le type de cancer) mais aussi l'ensemble des gènes dont les niveaux d'expression ont été affectés par ces fortes différences de conditions physiologiques. Bref, il sera relativement difficile d'identifier les gènes responsables ou impliqués dans un mécanisme précis.
Par ailleurs, quand il y a beaucoup de variations non contrôlées au cours de l'expérimentation, il devient difficile de distinguer des fluctuations aléatoires d'un effet spécifique [52]. Si l'on cherche à

identifier un nombre réduit de gènes responsables d'un mécanisme précis, il est avantageux d'obtenir des profils d'expression relativement proches pour la plupart des gènes. Il faut donc choisir des conditions qui ne font varier que très peu de facteurs et n'impliquent qu'un nombre réduit de gènes. Dans leur étude sur le mécanisme d'entrée et de métabolisme de deux sources de soufre, Sekowska *et al.* [52] ont choisi d'étudier le transcriptome dans des conditions très proches, à savoir la croissance des bactéries dans un milieu contrôlé en présence de méthionine ou de methylthioribose comme source de soufre (au plus une dizaine de gènes impliqués théoriquement). Malgré les conditions contrôlées, de nombreux gènes étaient détectés car le changement de conditions déclenchait leur voie métabolique. Ici la voie de la synthèse de l'arginine a été activée indirectement lors du changement de source de soufre. Les gènes de synthèse de l'arginine ne sont cependant pas impliqués directement dans l'assimilation du soufre. Par ailleurs, l'expérience a révélé le déclenchement involontaire de transitions dans les conditions environnementales avec les cascades de régulations qu'elles provoquent. Ces différences d'expression seraient dues à des différences de températures de la pièce pour la période comprise entre la phase préculture et la phase de croissance. Ainsi, même en choisissant des conditions proches, on retrouve d'autres mécanismes qui se mettent en place indépendamment du phénomène étudié.

De petites différences dans des conditions expérimentales non contrôlées peuvent mener à des différences d'expression de gènes visibles et cohérentes, concernant par exemple les gènes de compétence ou de sporulation. Si l'on avait choisi des conditions lointaines, il aurait été encore plus difficile de distinguer les gènes réellement impliqués dans le mécanisme d'intérêt des gènes annexes.

Deuxième étape : choix des autres facteurs pris en compte

Généralement, les facteurs autres que le facteur d'intérêt ne sont pas identifiés en tant que tels. On parle souvent de répétitions ou de réplications.

Les réplications ou répétitions des conditions expérimentales sont nécessaires afin de distinguer les variations d'expression présentes par hasard de celles reproductibles et réellement liées au facteur

d'intérêt. Sans réplication ou répétition il est impossible d'estimer l'erreur ou le bruit, paramètre nécessaire pour les méthodes d'analyse ultérieures [73].

Il est possible de distinguer deux niveaux de réplications :

- les répétitions techniques qui correspondent, par exemple, à deux spots du même gène sur la puce ou le même ARNm hybridé sur plusieurs puces ou encore différents protocoles d'extraction ou de marquage. Ces répétitions permettent d'évaluer la variabilité liée au protocole expérimental.
- la réplication vraie ou biologique comme l'ARNm de plusieurs individus, échantillons, spécimens. Cette réplication permet de prendre en compte la variabilité biologique. Les conclusions de l'analyse portent plus sur la population étudiée que sur l'échantillon en particulier. Elle comprend également des réplications de l'expérience à deux jours/ temps différents afin d'étudier la reproductibilité de l'expérience dans le temps. Les résultats obtenus sont donc plus fiables et reproductibles que si aucune réplication biologique n'avait été réalisée [77].

Par rapport aux réplications et aux répétitions, on peut trouver deux types de comportements erronés dans la mise en place des puces. Le premier est de craindre que les différences génétiques n'interfèrent dans les effets du facteur d'intérêt. Ainsi, certains expérimentateurs cherchent à homogénéiser les individus ou échantillons utilisés ou encore à sélectionner une lignée consanguine la plus adaptée au problème étudié. Or l'article de Turk *et al.* [78] montre que les variations d'expression entre deux lignées consanguines de souris sont relativement faibles. Ceci est rendu particulièrement évident par la comparaison du nombre de gènes différentiellement exprimés entre deux lignées ou entre un tissu atteint d'une maladie et un tissu sain. Ils en concluent que le profil génétique ne devrait interférer que marginalement dans l'analyse du transcriptome. Au contraire, Whitehead et Crawford [79] constatent que dans l'étude d'un groupe de gènes essentiels dans le métabolisme cellulaire (192 gènes), seuls 31% des gènes différentiellement exprimés entre deux tissus sont conservés entre des populations de poissons différentes. Ils en concluent que lorsque l'on fait les mesures sur une seule population, l'observation de différences très significatives d'expression dans certains tissus ne correspond pas nécessairement à des gènes représentatifs des différences

fonctionnelles ou morphologiques entre ces tissus. Les grandes variations d'expression entre individus [79] montrent qu'il est important d'inclure des replicats biologiques à l'intérieur des groupes de traitement afin d'attribuer des différences d'expression au traitement plutôt qu'à des variations entre individus ou populations. Les gènes détectés indépendamment de l'individu ou de la population expliqueront de manière plus probable les différences fonctionnelles et morphologiques étudiées.

Le deuxième comportement erroné est d'accorder beaucoup plus d'importance aux répétitions techniques qu'aux réplications biologiques. Souvent, on observe des plans d'expérience où il y a de nombreuses répétitions techniques mais peu de réplications biologiques. S'il est naturel d'avoir des résultats différents entre deux individus, on peut trouver plus inquiétant de ne pas avoir le même résultat pour un même échantillon selon le marquage ou la position sur la puce. La question se pose alors de la fiabilité et de la précision des mesures. Actuellement les mesures de puces sont globalement fiables dans la mesure où elles sont reproductibles au sein d'un même laboratoire. Il faut garder à l'esprit que tout comme la biologie, l'instrumentation subit l'influence de facteurs environnementaux et les variations techniques ne sont pas plus étonnantes que les variations biologiques. Si l'accent est mis sur les répétitions techniques au dépend des répétitions biologiques, les résultats obtenus sont alors précis et indépendants de la technique employée mais ne sont pas forcement généralisables à tout échantillon biologique similaire.

Pour conclure, lors de l'élaboration du plan expérimental, il est donc important de définir les différentes questions posées et de les hiérarchiser entre elles mais aussi de prendre en compte les contraintes matérielles comme le nombre d'hybridations possibles ainsi que les sources de variabilités techniques et biologiques [70]. On classe les facteurs de variations d'expression en trois catégories [70] :

- la variabilité biologique qui est intrinsèque à tous les organismes et dépend des facteurs génétiques et environnementaux. Elle est évaluée grâce aux différents réplicats
- la deuxième est technique (puce/extraction/ marquage/hybridation). Elle est évaluée grâce aux répétitions techniques
- la dernière est l'erreur de mesure.

Dans une expérience sur l'étude du métabolisme du soufre chez *B. subtilis* [52], les sources de variations les plus fortes sont la quantité d'ARNm utilisée pour synthétiser de l'ADNc qui influence plus certains gènes que d'autres (variabilité technique) et la difficulté de reproduire exactement les conditions expérimentales d'une expérience à une autre (jour de l'expérience, variation biologique). Les variations d'expression selon le facteur d'intérêt étaient faibles : cela était souhaité car les différences de conditions ne devaient faire intervenir qu'une dizaine de gènes.

L'agencement des différents facteurs : le plan d'expérience proprement dit

Prenons le cas où l'impact du facteur expérimental d'intérêt (mutation, conditions de culture, …) est confondu avec les variations causées par d'autres facteurs, par exemple deux jours de culture différents nécessités par le plan expérimental. Il est impossible de savoir si les variations d'un gène donné sont dues au facteur d'intérêt ou à un autre facteur [73]. Pour être plus clair, les résultats ne peuvent pas être interprétés correctement à cause du biais expérimental et l'ensemble de l'étude ne repond pas au but initial [71].

L'agencement des facteurs par rapport aux conditions d'expérience est donc primordial.

Revenons maintenant désormais à l'élaboration du plan expérimental. Il comprend différents points :

1. La définition de l'unité expérimentale : quel est le meilleur échantillon à utiliser pour réduire par exemple, la variance biologique : *pooler* ou non les échantillons. En théorie, rassembler les échantillons de plusieurs individus devrait augmenter la précision de la mesure en réduisant la variance des comparaisons d'intérêt [79]. Cependant, cette option présente l'inconvénient de potentiellement laisser un seul échantillon avoir une influence trop forte sur les résultats. Par ailleurs, elle ne permet pas d'estimer la variance biologique entre individus [79]. Le pooling s'avère parfois nécessaire notamment lorsque la quantité d'ARNm prélevée est faible.
2. Le nombre de répétitions. Actuellement, il n'existe pas de consensus sur le nombre de répétitions nécessaire. La réponse différerait selon l'expérience (la thématique, l'organisme, les conditions et autres). La tendance générale est : « plus on fait de répétitions, mieux c'est ». Il existe cependant

des méthodes pour définir *a posteriori* si le nombre d'échantillons biologiques est suffisant pour discriminer différents groupes cellulaires [80] par exemple différents types de leucémies. Mais comme ces méthodes sont réalisées *a posteriori*, elles ne sont pas de grande utilité pour définir un plan expérimental.

3. La manière d'associer les échantillons. Les plans d'expérience complets et réguliers sont ceux qui permettent les analyses les plus puissantes. Sinon, pour un nombre donné de puces, les plans d'expérience équilibrés pour les facteurs d'intérêt sont les plus efficaces [73]. Il n'est pas toujours facile de les mettre en place, car soit ils requièrent un grand nombre de mesures, ce qui pose des problèmes de coût, soit parce qu'un individu –chez l'Homme par exemple– ne peut présenter les différents modes du facteur d'intérêt. De manière caricaturale, un individu peut être difficilement à la fois malade et sain. Plusieurs plans expérimentaux sont décrits dans Kerr *et al.* [81] dans le cadre d'une puce à ADN avec hybridation simultanée de deux échantillons, un marqué en rouge, l'autre en vert :
    - le carré latin communément appelé dans le domaine *dye swap* : les échantillons d'ARNm sont marqués par deux marqueurs différents une fois rouge et une fois vert. Ce plan d'expérience permet de mesurer les biais du marquage sur les mesures de transcriptome.
    - un autre plan communément utilisé est l'utilisation d'une mesure de référence. Les échantillons d'intérêt sont toujours marqués de la même couleur et l'échantillon de référence de l'autre. Ce plan d'expérience ne permet donc pas l'estimation de l'effet fluorochrome (une répétition technique). En outre, avec ce plan expérimental on obtient beaucoup plus de mesures de l'échantillon de référence que des échantillons d'intérêt. Il y a donc une perte d'argent et de temps en faisant des mesures « inutiles » [73],[82].
    - utiliser à la place un plan équilibré permet d'acquérir deux fois plus de données pour autant de moyens [73]. Les mesures ainsi effectuées sont donc beaucoup plus précises. Dans un plan équilibré, chaque mode de facteur expérimental est répété le même nombre de fois. Le plan expérimental équilibré de l'article de Sekowska *et al.* [52] peut être une base pour mettre au point un plan expérimental adapté à la problématique choisie.

Le choix de la puce ou la préparation de la puce

Si une des premières mesures d'expression [17] a utilisé une puce contenant 45 ADNc d'*Arabidopsis thaliana*, les supports actuels permettent la mesure simultanée de l'expression de plusieurs dizaines de milliers de gènes. Désormais, il existe une grande variété de puces en adéquation avec l'objectif de l'expérience. Certaines de ces puces sont génomiques et représentent l'ensemble du génome de l'organisme, d'autres sont dédiées à quelques gènes généralement impliqués dans un même processus cellulaire. En plus de la diversité des gènes étudiés, il existe différents types de supports, de sondes, de densités et de marquages de la cible utilisés [3].

Le choix de la puce est plus ou moins ample suivant qu'on utilise des puces commerciales ou que l'on produise la puce. La description suivante précise les différents types de puces existants ainsi que leurs qualités respectives.

Historiquement, les *macroarray*, les *microarrays* et les « véritables » puces à ADN correspondent à trois techniques différentes [3].

- Les *macroarrays* utilisent comme sonde des clones d'ADN complémentaire (ADNc) disposés sur des membranes de nylon avec un espacement de l'ordre du millimètre en association avec des cibles radioactives. Les sondes représentent entre 200 et 8 000 gènes. Leur densité est donc relativement faible. Cette technique ne demande pas d'équipement particulier à part le *phosphorimager*.
- Les *microarrays* plus miniaturisés, comportent quelques milliers de gènes représentés par des produits PCR déposés tous les 200 à 400 microns sur une lame de verre et des cibles marquées par fluorescence. Cette technique permet l'hybridation compétitive.
- Les « véritables » puces à ADN associaient à chacun des gènes d'un organisme un ensemble de sondes sous la forme d'oligonucléotides synthétisés *in situ* sur la surface de la matrice. Les oligonucléotides mesurent au plus vingt-cinq bases à cause de l'efficacité finie de chaque étape [74]. Chaque puce peut comporter jusqu'à 40 000 à un million d'oligonucléotides. Afin de mesurer la possibilité d'hybridation croisée, certains

oligonucléotides correspondent à la séquence exacte et d'autres comportent une mutation « mismatch » au milieu.

Aujourd'hui, ces trois distinctions n'ont plus vraiment lieu d'être, d'autant plus que ces techniques sont utilisées de façon croisée comme le montre l'exemple de puces à ADN utilisant des produits PCR et des cibles radioactives. Globalement, il est encore possible de distinguer les *microarrays* des *macroarrays* du fait de la densité des sondes sur la puce. Les terminologies « puces à ADN » et « *microarray* » sont donc employées de façon indifférente. Les termes « biopuces » ou « microréseau » sont également employés.

Comme il est difficile de définir chaque type de puce précisément, nous aborderons plutôt leurs différents composants : le support, le type de sonde, le type de marquage.

Le type de sonde

Les sondes ADNc

L'utilisation de sondes ADNc ne nécessite pas la séquence complète du génome mais l'utilisation de banques d'ADNc issus des différents EST (Expressed Sequence Tag) disponibles. Les banques d'ADNc correspondent à des inserts dans des plasmides ou des chromosomes bactériens [83]. L'ADNc est ensuite récupéré grâce à l'extraction et la purification des inserts puis à l'amplification PCR avec des primer universels. Les produits sont ensuite séparés sur gel d'électrophorèse, quantifiés et déposés sur le support [74].
Ce type de sonde est donc recommandé pour les organismes dont le génome n'est pas séquencé. Il est cependant également largement utilisé pour l'ensemble des organismes (dont le génome est séquencé ou non) car il ne nécessite pas de matériel spécifique comme les oligonucléotides.

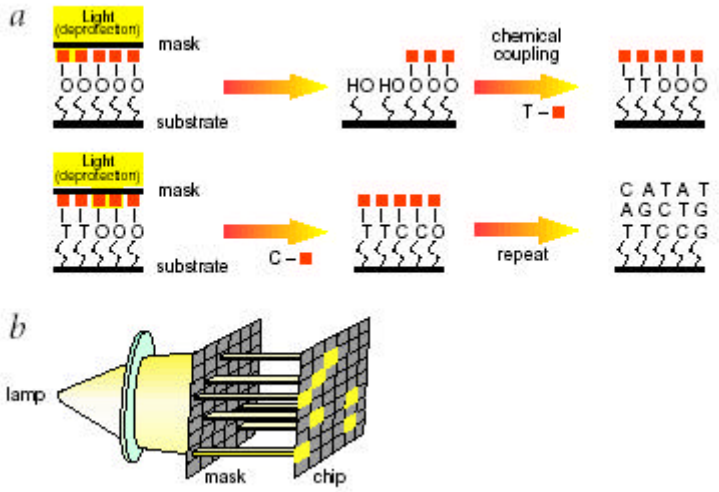
Ces sondes sont relativement longues puisqu'elles comprennent, en moyenne, entre une centaine et cinq cents paires de bases. Cette particularité limite les possibilités d'hybridation croisée sauf dans le cas de séquences très proches comme certaines familles de gènes très similaires chez les plantes ou

pour des isoformes d'une même famille [84]. La modification du design des puces est par contre une chose aisée puisqu'il suffit de rajouter un spot contenant l'ADNc qui correspond au gène à étudier.

Une des difficultés d'utiliser des ADNc est le maintien des banques d'ADNc qui peut s'avérer coûteux et lourd pour un laboratoire isolé [74]. Ces banques contiennent entre 1 et 5% de séquences redondantes, mal annotées ou encore contaminées [83]. Certaines banques d'ADNc commerciales sont désormais disponibles pour les organismes d'intérêt tels que l'homme, la souris, le rat et le chien. Par ailleurs Lipshutz *et al.* [25] pointaient le risque de mauvaise identification du spot ou de l'ADNc déposé et du coup une mauvaise attribution du niveau d'expression. Ce risque a été confirmé par Kothapalli *et al.* [84] qui a re-séquencé 17 ADNc qui correspondent à des gènes différentiellement exprimés dans leur expérience. Parmi ces dix-sept ADNc, quatre (24%) présentent des séquences incorrectes qui ne correspondent pas au gène qu'ils devaient représenter.

Les oligonucléotides courts

Les puces à oligonucléotides sont fondées en majorité sur une technique de fabrication de puces informatiques [20] adaptée ensuite par la société Affymetrix à l'étude de l'expression des gènes [18, 25]. Les sondes sont des oligonucléotides courts composés de 25 paires de bases synthétisées *in situ* par des dépôts de couches successives de quatre nucléotides. La technique débute par l'attache à la surface de la puce de liaisons synthétiques munie de groupes chimiques qui peuvent être enlevés sous l'effet de la lumière. On envoie ensuite de la lumière à certaines localisations définies *via* un masque photolithographique afin de déprotéger les liants. Le rajout de deoxynucleosides avec un groupe photolabile permet la liaison avec les groupes déprotégés. Un autre masque dont la configuration varie pour chaque couche déposée et qui assure ainsi une succession correcte des bases est ensuite mis en place et ainsi de suite jusqu'à la synthèse complète des oligonucléotides. La Figure 4 illustre cette technique.



a Synthèse d'oligonucléotides par photo-activation. Des molécules terminées par un groupe protecteur photolabile sont fixées sur un support solide. De la lumière est dirigée à travers un masque afin de déprotéger et d'activer les sites sélectionnés. Des nucléotides protégés s'assemblent alors avec les sites activés. Le processus est répété : activation de différents sites puis assemblage des différentes bases. Ce procédé permet la construction de sondes ADN spécifiques sur chaque site.
b : représentation schématique de la lampe, du masque et de la puce

Figure 4 : Technique de synthèse des oligonucléotides courts (figure tirée de [25])

Toutefois, la technique de l'impression jet d'encre [85] est également utilisée afin de fabriquer les oligonucléotides *in situ*. Le principe est simple : une aiguille de l'imprimante permet d'appliquer une seule goutte de dix picolitres à une position identifiée. Les étapes sont les mêmes que la synthèse par photolithographie : protection, déprotection, synthèse. A la place de l'action de la lumière, des agents chimiques permettent la réalisation des réactions qui conduisent à la synthèse des oligonucléotides.

Pour les puces Affymetrix, l'expression de chaque gène est mesurée par une vingtaine d'oligonucléotides différents [74]. Afin d'identifier d'éventuelles hybridations croisées et éliminer le bruit résultant, chaque portion du gène choisie est représentée par des oligonucléotides qui correspondent à la séquence exacte de la cible (PM ou Perfect Match) et par d'autres dont la séquence est identique sauf une mutation située à la position centrale (MM : Mismatch) (Figure 5).

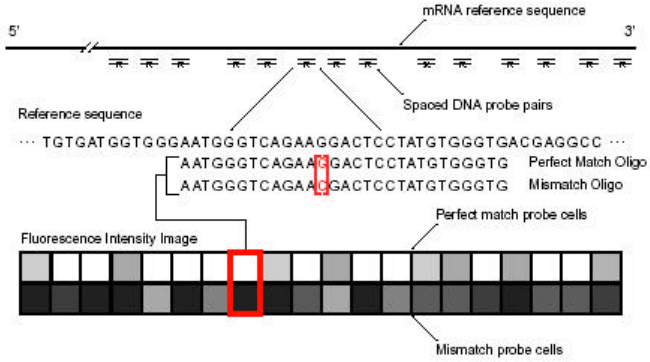


Figure 5 : Exemple d'oligonucléotides PM et MM (image tirée de [25])

Toutes les sondes sont de même longueur et présentent, autant que faire se peut, la même composition en G/C ce qui leur procure l'avantage d'avoir une température d'hybridation commune ou presque, contrairement aux ADNc. Par ailleurs, chaque spot contient la même quantité d'oligonucléotides contrairement, là encore, aux ADNc [74].

Les puces à oligonucléotides sont préférées pour l'analyse complète des génomes [19]. La possibilité de représenter toute séquence présente dans un génome, la petite longueur de ces sondes et la possibilité de sondes chevauchantes multiples permet de détecter des caractéristiques génomiques comme de petits polymorphismes, des variants d'épissage, la distinction entre des membres d'une famille génique ou la distinction de régions répétitives.

Les oligonucléotides présentent l'avantage de ne pas à avoir à entretenir une banque d'ADNc et réduit le risque d'avoir des spots mal identifiés [25]. Chaque gène est représenté par plusieurs sondes, soit plusieurs mesures pour un même gène.

La principale difficulté selon Lipshutz *et al.* [25] est la sélection des oligonucléotides dont la séquence doit être spécifique du gène et si possible non chevauchante avec les autres oligonucléotides chargés de détecter le même gène. Par ailleurs la synthèse *in situ* limite la possibilité d'adapter les puces aux différentes expériences. Il est difficile de rajouter de nouvelles sondes pour de nouveaux gènes ou portion d'ADN à identifier puisqu'il est alors nécessaire de refabriquer l'ensemble des masques

photolithographiques. Par ailleurs, les chercheurs n'ont pas accès à la séquence des sondes utilisées et doivent travailler sur les annotations fournies par le fabricant [86].

Les oligonucléotides longs

Ces sondes, comme leur nom l'indique, sont plus longues que les oligonucléotides courts puisqu'elles se composent de 40 à 80 paires de bases. Contrairement aux oligo-nucléotides courts, leur synthèse ne s'effectue pas *in situ.* Ils sont généralement déposés par impression jet d'encre [85]. En 2001 Hughes *et al.* [87] ont élargi cette technique initialement utilisée pour la fabrication des oligonucléotides courts à celle des oligonucléotides longs.

Les oligonucléotides longs comportent les même avantages que les oligonucléotides courts avec, en plus, la possibilité d'identifier les différents transcrits alternatifs [74, 88]. Ils présentent moins de risques d'hybridation croisée que les oligonucléotides courts. Cependant Kane *et al.* [88] précisent que si un ARNm présente plus de 70% d'identité avec l'oligonucléotide long, il y a de grands risques d'hybridation croisée.

La Figure 6 tirée de l'article de Barrett et Kawaski [74] résume les différents types de sondes disponibles et leurs différents modes d'obtention.

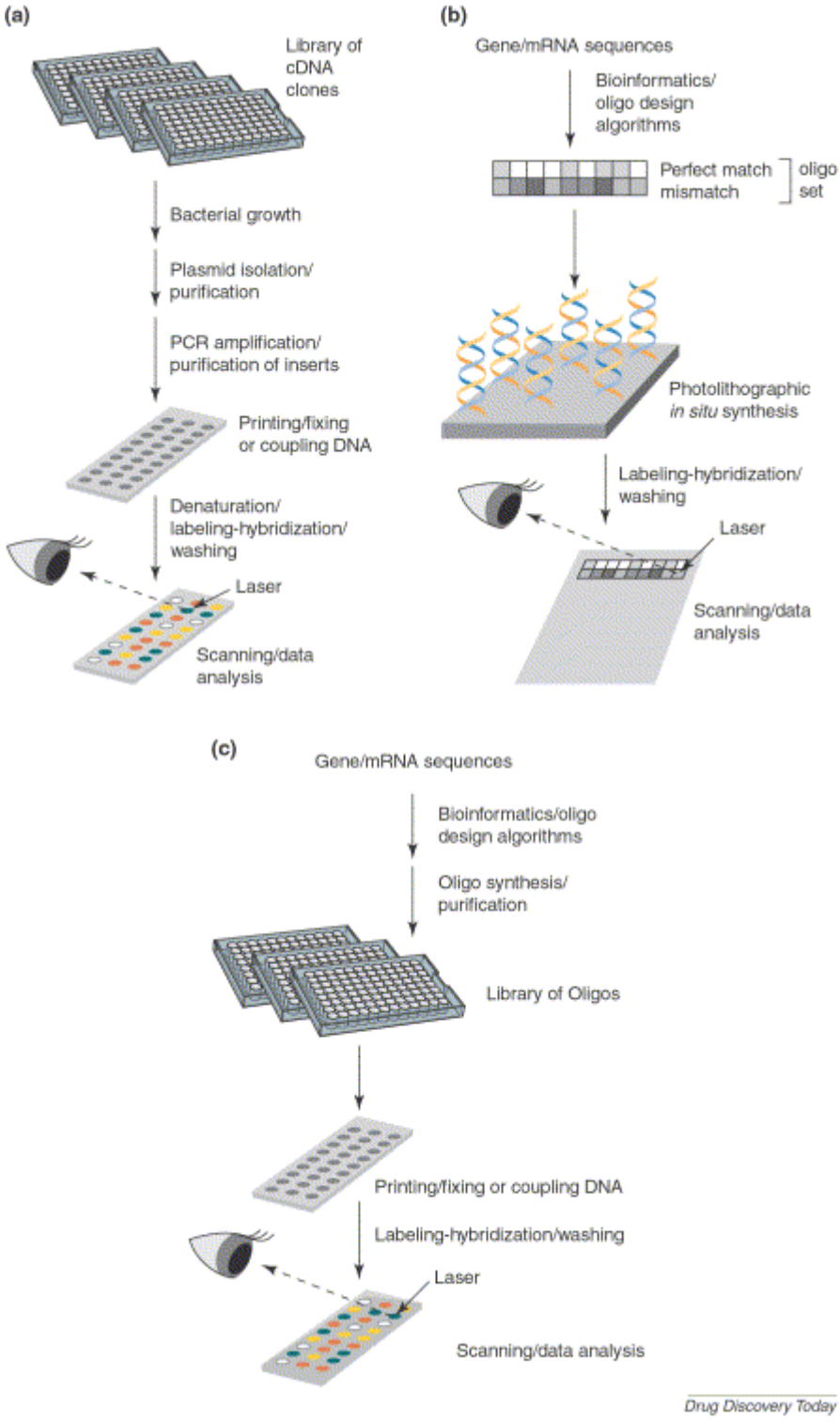


Figure 6 : Les différents types de sondes disponibles

Comparaison des différentes sondes

Le problème principal des oligonucléotides est leur conception. Il est nécessaire d'identifier des séquences spécifiques d'un gène, peu semblables à d'autres gènes, situées entièrement dans un exon, sans séquence répétitive ni la possibilité de palindrome et avec une composition en G/C constante pour toutes les sondes [74].

Kane *et al.* [88] ont comparé la différence de sensibilité entre des oligonucléotides longs (50 bp) et des produits de PCR (322 à 393 pb). Ces deux types de sondes présentent une sensibilité comparable. Leurs seuils de détection seraient de dix copies d'ARNm par cellules.

Des études ont comparé les résultats obtenus à partir de sondes différentes. La plupart montrent que les corrélations entre les résultats obtenus avec deux sondes différentes sont relativement faibles [89]. Mais la suite de leurs résultats ne sont pas concordants. Certains précisent que les oligonucléotides courts donnent des résultats plus fiables que les ADNc, d'autres démontrent l'inverse. Enfin Yuen *et al.* [90] précisent que les oligonucléotides et les ADNc conduisent à une sensibilité et une spécificité équivalente.

Par ailleurs, l'incohérence des résultats se retrouve également pour le même type de sondes (ici des oligonucléotides) mais sur deux puces différentes. Nimgaonkar *et al.* [91] ont montré que deux générations de puces Affymetrix pour l'Homme donnent des résultats relativement incohérents du fait d'un changement de densité, de sélection des oligonucléotides et d'autres changements dans la mise au point de la puce. Ces résultats ont été confirmés ultérieurement avec deux générations de puces pour l'étude du cancer [92]. Pour 25% des gènes, les résultats sont incohérents entre les deux types de puces.

Le type de support

Deux types de supports principaux sont actuellement utilisés : les membranes et les lames de verre.

Les filtres ou membranes de nitrocellulose sont utilisées généralement avec un marquage radioactif. Ils peuvent être utilisés plusieurs fois, contrairement au support de verre. Les membranes avec marquage radioactif ont montré une plus grande sensibilité que les lame de verre avec fluorescence [93].

Les lames de verre sont le support favori car elles ont une fluorescence résultante faible, elles sont transparentes et résistantes aux hautes températures. Comme le liquide ne peut pas pénétrer dans la lame, la surface, les sondes et les cibles sont directement en contact sans diffusion *via* des pores. Bien sûr, cela demande d'agiter la préparation afin d'obtenir un haut taux d'hybridation (ceci est valable également pour les membranes) [83, 94]. Contrairement aux membranes de nylon, leur relative « planité » autorise une lecture avec une précision de dizaine de micromètres. Leur rigidité permet une meilleure localisation des différents spots et plusieurs modifications chimiques de la surface sont disponibles afin de fixer les sondes [83, 94].

Toutefois Stillman et Tonkinson [93] ont montré que la longueur de la sonde avait une plus grande influence sur la qualité de l'hybridation que le type de support. Cependant, les lames de verre sont le support privilégié sans doute à cause de la possibilité de réaliser l'hybridation de deux échantillons à la fois et ainsi comparer directement deux conditions expérimentales.

Le type de marquage

La plupart des puces actuelles utilisent un marquage fluorescent avec deux marqueurs de longueurs d'ondes d'émission différentes (une dans le vert et l'autre dans le rouge). Les cyanines fluorescentes Cy3 (vert) et Cy5 (rouge) mais aussi la fluorescéine et la rhodamine sont le plus souvent utilisées. La double fluorescence permet l'étude simultanée de deux échantillons sur la même puce. Cette possibilité explique l'engouement pour ce type de marquage puisqu'ainsi les biais entre différentes puces sont éliminés. Cependant 't Hoen *et al.* [95] ont montré que les valeurs obtenues ne sont pas influencées par la présence d'un seul ou de deux fluorochores. Ils conseillent, par ailleurs, d'utiliser les valeurs obtenues pour chacun de ces marquages plutôt que le ratio entre ces valeurs généralement utilisé. Aussi des marquages avec simple fluorescence sont également employés.

Une solution alternative est le marquage radioactif au ^{32}P ou ^{33}P qui ne nécessite que de petites quantités d'ARN [83]. La lecture se fait à l'aide d'un *phosphorimager* couramment disponible au sein de laboratoires de biologie. Querec *et al.* [96] ont montré que la radioactivité présente une plus grande sensibilité par rapport à la fluorescence et une reproductibilité supérieure. Cette plus grande sensibilité résulte de la nature du marquage radioactif : à des temps d'exposition suffisamment longs, l'émulsion sensible aux rayons X sera « activée ». Cependant, le temps d'exposition pour maximiser la détection des signaux faibles peut gêner la détection des gènes les plus exprimés à cause d'une saturation du signal. Les radiations excèdent alors la limite de détection du film ou du *phosphorimager.*

L'extraction des ARNm et le marquage

L'extraction est une des étapes clés de la mesure du transcriptome. Vu le temps de demi-vie variable des différents ARNm (renouvellement des ARNm de l'ordre d'une à deux minutes pour *E. coli* et *B. subtilis*), une extraction rapide est préférée. Il est également nécessaire de limiter le stress de l'extraction afin de limiter les perturbations du système et la synthèse d'ARNm des protéines de choc thermique. Les protocoles varient selon les organismes étudiés et leurs stades. Les mesures d'expression dépendent fortement de la technique d'extraction utilisée et de la sensibilité spécifique de chaque gène à la dégradation de son ARNm [97].

Il est parfois nécessaire d'amplifier les ARNm obtenus, du fait du trop faible nombre d'exemplaires présents. Cette étape d'amplification PCR implique malheureusement des biais selon les gènes. Aussi, il est préférable de l'éviter si elle n'est pas nécessaire.

Après extraction des ARNm, il est préférable de synthétiser les ADNc correspondants, plus stables. Pour cela, il est possible, chez les eucaryotes, d'utiliser la queue polyA présente à l'extrémité 3' des ARNm pour ancrer une amorce polyT qui permet, par une transcriptase inverse, la synthèse d'ADNc marqués. Une autre possibilité lorsque la queue polyA est inexistante est la synthèse d'ADNc à partir d'un ensemble d'amorces aléatoires ou spécifiques des ARN à détecter. La longueur de l'ADNc synthétisé peut varier, ce qui influence également les niveaux d'hybridation mesurés [97]. Par ailleurs, la quantité d'ADNc synthétisée n'est pas proportionnelle à la quantité d'ARN utilisée et pire, le

rendement de la synthèse d'ADNc dépend d'un gène à un autre [52]. Une des hypothèses est que, pour certains gènes, la synthèse d'ADNc peut être affectée par la formation de structures secondaires dans l'ARNm ou par la présence de segments d'ARN largement biaisés par la composition en nucléotide [52].

Afin de révéler les ADN présents dans la cellule, une molécule radioactive ou fluorescente est incorporée (cf. types de puces). Cependant, l'incorporation de ces molécules peut varier selon l'ARNm. Ainsi la composition en nucléotides de l'ADNc fait varier la qualité du marquage radioactif [97].

L'hybridation

La composition des sondes influence la stabilité de l'hybridation dans des conditions fixées. Ces impacts ont été relativement bien étudiés de façon à déterminer les conditions optimales d'hybridation. Le taux de GC, la structure des acides nucléiques ou la localisation de la sonde sur la séquence du gène influencent la température et les solutions nécessaires à l'hybridation. On peut se référer à l'article de Maskos et Southern pour l'étude des conditions pour les oligonucléotides [98].

La lecture des données

La phase de lecture de données est loin d'être anodine. L'acquisition des images conditionne de façon majeure la précision des données et donc la pertinence des interprétations [75]. Suivant le type de marquage, l'acquisition des images diffèrent. Afin de détecter la fluorescence émise par la puce, les fluorophores sont excités à leur fréquence par un laser tandis qu'un scanner ou un microscope confocal couplé à un tube photomultiplicateur (PMT) permet l'analyse des photons émis par les marqueurs. Les canaux de lecture correspondant aux longueurs d'onde 635 nm et 532 nm sont utilisés pour lire la fluorescence de Cy5 et Cy3. Dans le cas d'un marquage radioactif, la radioactivité est révélée par autoradiographie grâce à une exposition d'un film à rayon X ou à un scanner

phosphorimager qui permet de détecter la radioactivité présente pour chaque spot. Une image est alors obtenue pour chaque échantillon (2 images pour les puces avec deux fluorochromes).

Les paramètres de réglages des appareils influencent grandement la qualité des images obtenues. En effet les réglages des photomultiplicateurs pour la fluorescence et le temps d'exposition pour la radioactivité influent grandement sur la sensibilité de la puce.

Ainsi pour la radioactivité, si de grands temps d'exposition sont utilisés, il sera possible de détecter des ARNm présents à des niveaux très faibles. Cependant, on s'expose alors à des problèmes de saturation du film pour les gènes fortement exprimés ainsi. A cela s'ajoute le fait que l'intensité élevée d'un spot aura tendance à masquer l'intensité des spots voisins. Au contraire, des temps d'exposition courts écarteront tout problème de saturation et permettront une bonne évaluation de l'expression des gènes fortement exprimés. Cependant, les ARNm peu présents dans la cellule ne seront sans doute pas détectés. Aussi, utiliser différents temps d'exposition permet d'obtenir une mesure correcte, une grande gamme d'intensités et d'augmenter le nombre de gènes détectés [96].

Pour la fluorométrie, les scanners disposent généralement d'options diverses qui permettent d'améliorer la qualité du signal détecté : plus on augmente la tension du PMT ou la puissance du laser, plus les intensités mesurées sont fortes. Généralement, les intensités faibles sont détectées avec une précision plus faible. Il est donc tentant d'ajuster les paramètres afin d'obtenir des intensités fortes même pour les spots faiblement marqués. L'ajustement le plus couramment utilisé consiste à régler le gain du PMT de façon à ne conserver qu'un nombre minimal de spots présentant des pixel saturés [2]. La puissance du laser est plus rarement utilisée du fait du risque de blanchissement de l'image (*photo-bleaching*) à trop fortes puissances [75]. Cependant, des erreurs de mesure significatives peuvent être introduites lors de lectures à différentes tensions [99, 100].

Généralement, le signal mesuré est supposé proportionnel à l'intensité de la lumière émise à un voltage donné. Lyng *et al.* [99] ont rappelé pour trois scanners l'influence non linéaire de la tension PMT sur l'intensité moyenne mesurée. La plupart des scanners montrent la relation log-linéaire désirée pour les moyennes d'intensités d'expression comprises entre 200 et 50 000 (Figure 7). Au-

dessus, il y a saturation, l'intensité croit plus lentement puis se stabilise. En dessous, l'intensité décroît rapidement avec un taux de décroissance différent selon les spots. Des problèmes de discrétisation des variables peuvent surgir pour des valeurs inférieures à 200. Lyng *et al.* [99] préconisent de faire différents scans pour des tensions différentes à partir d'une même puce en fonction de l'intensité des spots les plus faibles/forts.

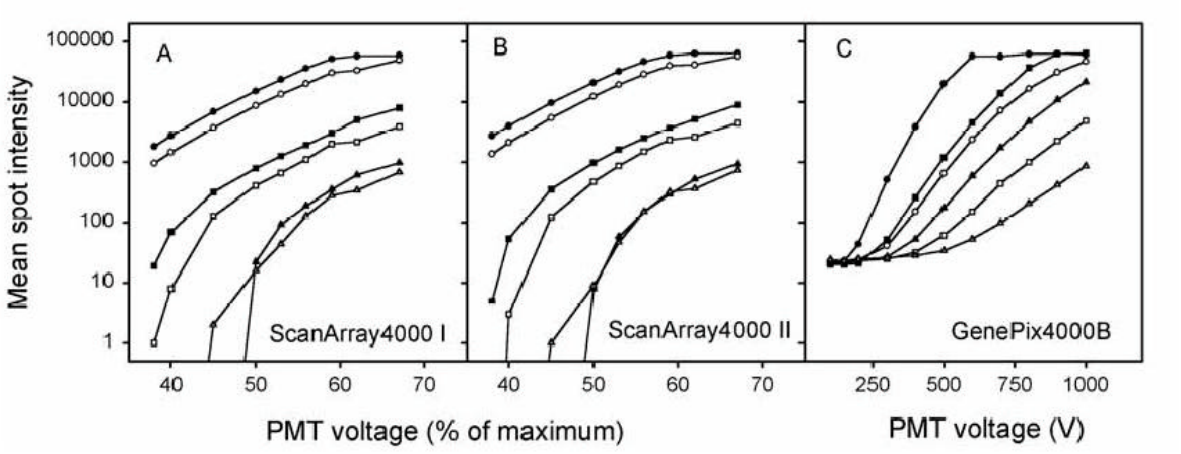


Figure 1
Mean spot intensity *versus* PMT voltage for three scanners of two different brands, ScanArray4000 I (A), ScanArray4000 II (B), and GenePix4000B (C). Three representative spots are presented, one with high intensities (●, ○), one with intermediate intensities (■, □), and one with low intensities (▲, △). The intensities of the green and red channel are represented by closed and open symbols, respectively. The PMT voltage is given in % of maximum voltage for ScanArray4000 and in voltage for GenePix4000B.

Figure 7 : Gamme de linéarité pour la mesure de niveaux d'expression (figure tirée de Lyng *et al.* [99])

A partir de l'image obtenue, les différents spots sont reconnus *via* un logiciel à l'aide d'une grille [75]. Les principaux logiciels utilisés sont ScanAlyze [101], ImaGene et GeneSight de Biodiscovery inc, AtlasImage et AtlasNavigator de BD Biosciences Clontech et ArrayExplorer [102]. Généralement, pour la mise en place de la grille, certains paramètres de la puce sont précisés comme la disposition de la puce, la taille et la forme des spots. Au départ, on peut croire que nous connaissons la disposition des spots et leur écartement et qu'ils sont tous circulaires avec le même diamètre. Cette vision idéale se révèle erronée: beaucoup de spots ne se trouvent pas dans la grille exactement à l'endroit attendu et certains présentent une forme non circulaire. Par conséquent, des interventions humaines sont souvent nécessaires à cette étape afin de correctement repositionner les spots ou affiner leur forme et leur taille [76]. Aussi, deux chercheurs qui utilisent le même logiciel, par exemple scanAlyze, pour traiter la même image ne trouvent pas forcément des résultats identiques ou cohérents [103]. Cela

peut entraîner des variations fortes entre deux mesures, comme des ratios doublés simplement à cause du placement de la grille.

L'intensité des spots est ensuite extraite. Globalement, le niveau d'expression d'un gène correspond à la moyenne ou la médiane des intensités des pixels du spot qu'il représente.

Pour conclure, Stoyanova *et al.* [100] montrent que la plupart des effets non linéaires trouvés dans leur analyse correspondent à de mauvais réglages du scanner. Une attention toute particulière doit donc être apporté à ce paramètre.

La normalisation et la prise en compte du bruit

Le terme normalisation se réfère aux différents moyens d'éliminer l'effet des sources de variations systématiques qui affectent les mesures de transcriptome [104]. Cette étape peut s'avérer nécessaire afin de distinguer les différences d'expression biologiques des différences liées au protocole expérimental utilisé. Les polémiques les plus importantes sur la normalisation concernent la soustraction du signal par un signal de référence ou la correction du bruit de fond ainsi que les techniques de lissage [82].

Prise en compte du bruit

La prise en compte du bruit est généralisée dans la plupart des laboratoires. L'idée est qu'il existe une fluorescence résiduelle ou des hybridations non complémentaires dont il faut tenir compte afin d'obtenir le « véritable » niveau d'expression d'un gène. Cette prise en compte dépend du type de sondes utilisées.

La prise en compte du bruit pour les puces à oligonucléotides courts Affymetrix est réalisée à partir de la détection d'un signal sur des sondes mutées (MM). Il était au début préconisé de soustraire tout simplement le signal de la sonde mutée (MM) du signal de la sonde exacte (PM). Toutefois, cette

méthode présente l'inconvénient d'obtenir de nombreux signaux négatifs : la sonde mutée hybride quelques fois plus que la sonde exacte. Par ailleurs, cette soustraction est inadéquate pour les intensités fortes ou faibles [105]. Ainsi, pour les intensités fortes, le gène à détecter s'hybride sur les PM mais aussi sur les MM de manière non négligeable. En conséquence, les valeurs PM-MM deviennent faibles comparées à l'intensité d'expression réelle. Cet aspect du problème est décrit par Kothapalli *et al.* [84] : parfois le signal MM masque le signal PM. Ce qui fait qu'un gène peut être considéré comme non exprimé dans une condition alors qu'il l'est après vérification par d'autres techniques. En bref, la soustraction du signal MM au PM n'est pas justifiée et ce d'autant plus que la température d'hybridation actuelle dans le protocole expérimental est plus basse que la température qui, idéalement, permettrait l'hybridation des cibles avec le PM et non le MM [106].

Pour les ADNc, généralement, l'intensité résiduelle autour de chaque spot est évaluée et soustraite à l'intensité du spot. Là encore, les données résultantes comprennent des valeurs négatives. Par ailleurs, outre les questions de la méthodologie utilisée afin de déterminer les alentours du spot, d'autres problèmes sont soulevés. Dans la discussion de leur article, Lyng *et al.* [99] remarquent qu'en raison de l'étape d'acquisition de l'image en fluorométrie, il y a des erreurs importantes pour toutes les intensités faibles. Il en découle que la mesure du bruit est elle-même bruitée. Toutes les mesures dans lesquelles le bruit est pris en compte finissent par être biaisées à cause de l'évaluation du bruit. L'impact sera relativement faible pour les valeurs fortes mais il sera beaucoup plus important sur les valeurs faibles.

Pour être complet, il convient de se poser également la question du type de bruit de fond pris en compte. Pour les puces Affymetrix, le bruit que l'on tente de mesurer correspond à l'hybridation croisée entre des séquences homologues, il dépend donc du gène étudié. En revanche, pour les sondes ADNc, le bruit correspond à une hybridation non spécifique sur la surface du support [3]. La prise en compte du bruit n'est utile dans ce cas que si l'hybridation fluctue selon la localisation sur le support. Par ailleurs Vrana *et al.* [3] soutiennent que ce bruit de fond n'est pas le même au sein d'un spot (présence de sondes) ou à coté d'un spot (hybridation directe sur le support). Ils suggèrent de considérer le bruit comme l'intensité du spot le plus faible correspondant à une fixation non spécifique. Cependant, la soustraction du même bruit de fond pour l'ensemble des gènes risque de se révéler

inutile. Enfin, d'autres approches plus élaborées se sont développées comme une approche bayésienne qui utilise les mesures du bruit autour du spot [107]. Ces approches élégantes ne résolvent pas les problèmes liés au manque de précision dans la mesure de ce bruit ni la pertinence de sa prise en compte.

Pour conclure, la prise en compte du bruit est fortement sujette à débat. Ainsi, la correction du bruit basée sur la médiane de l'intensité des pixels autour du spot n'a pas été prouvée comme bénéfique [82]. La Figure 8 montre que la prise en compte du bruit risque de rajouter du bruit aux données initiales. Deux mesures du même échantillon marqué en radioactivité sont confrontées. A gauche, sont représentées les mesures sans prise en compte du bruit, à droite celles dont on a soustrait le bruit estimé à partir d'hybridations non spécifiques. Ce graphique démontre clairement le rajout de bruit pour les valeurs faibles en tentant de prendre en compte le problème d'hybridation non spécifique.

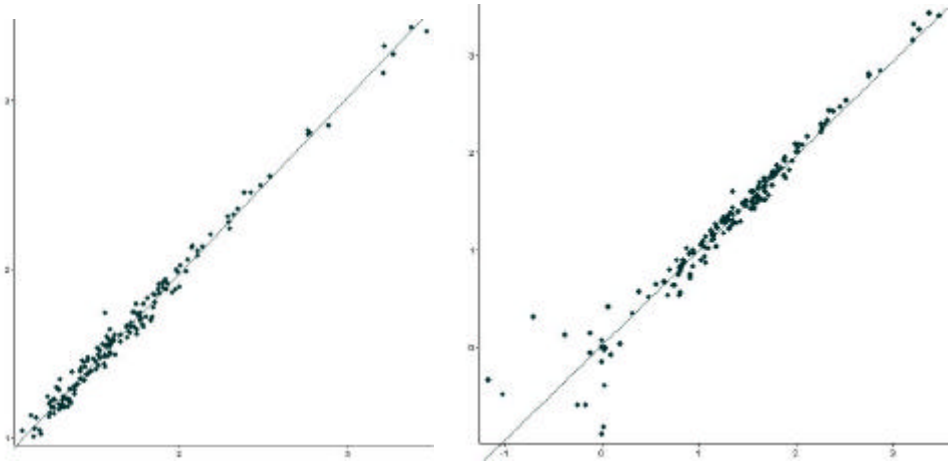


Figure 8 : La prise en compte du bruit rajoute du bruit

Différentes normalisations

Les sources de variations sont multiples : de l'effet des différents fluorochromes à l'effet du réglage du scanner. Généralement, afin de pouvoir comparer les mesures effectuées sur deux puces ou avec deux fluorochromes différents, il est utile de remettre à la même échelle les mesures et d'éliminer les biais selon les intensités d'expression [108]. Plusieurs types de normalisation peuvent être distingués.

Normalisation afin de limiter les effets de facteurs extérieurs comme le marquage et les réglages du scanner

Une des premières stratégies a consisté à normaliser les données à partir des mesures de gènes de ménages (*housekeeping gene*) dont le niveau d'expression est supposé invariant quelles que soient les conditions [3, 79, 100]. Il a été cependant révélé que leurs niveaux d'expression varient quand même pour certaines conditions expérimentales ce qui rend toute normalisation fondée sur leur expression aléatoire [109-111]. Depuis, il existe un lot de méthodes qui identifient un jeu de gènes supposés de ménage dans les données [100], [111]. Ces méthodes présentent l'avantage de ne pas faire d'hypothèse sur les variations d'expression réelles. Elles peuvent être intéressantes, notamment sur les puces dédiées. Cependant, elles dépendent fortement du lot de gènes identifié et requièrent un certain nombre de puces pour être valides.

Certaines expériences comparent plus de deux conditions expérimentales. Avec des puces à marquage fluorométrique, une référence marquée d'une couleur est généralement utilisée lors de chaque hybridation. La mesure du signal d'intérêt est parfois « normalisée » en soustrayant le signal de référence. Thygesen *et al.* [82] ont analysé l'efficacité de cette soustraction. Pour leurs données elle n'apporte rien, voire pire, elle dégrade le signal initial.

La plupart des méthodes de normalisation globales se fondent sur la moyenne des mesures de l'ensemble des gènes [100]. L'hypothèse sous-jacente suppose que l'expression de la majorité des gènes ne varie pas. Par ailleurs, les variations d'expression des gènes restants devraient s'équilibrer : autant de gènes seraient sous et sur-exprimés dans des échelles de variation proches [104, 108]. Enfin, entre deux conditions expérimentales, les niveaux moyens d'expression restent identiques (pas plus d'ARNm en moyenne). Une des normalisations les plus simples est de centrer et réduire l'ensemble des valeurs pour chaque condition. L'utilisation de la médiane et des quantiles au lieu respectivement de la moyenne et de la variance permet de se libérer de l'influence d'éventuelles valeurs extrêmes [3, 100]. Enfin Schuchhardt *et al.* [97] ont divisé les intensités par la moyenne de chaque condition expérimentale. Dans certaines conditions, ce type de normalisation est totalement

inadéquat. Ainsi, si l'étude porte sur la stabilité des ARNm au cours du temps, le niveau d'expression moyen va diminuer au fil des prélèvements du fait de la destruction des ARNm. Pour les puces dédiées à une thématique, où les gènes sont sélectionnés selon leur fonction, les hypothèses du maintien du niveau d'expression global et d'un faible nombre de gènes différentiellement exprimés peuvent également être invalidées [108, 112]. Par ailleurs, cette normalisation globale ne prend pas en compte d'éventuels bruits qui dépendent des intensités mesurées [111].

Kerr *et al.* [81] ont introduit l'ANOVA (*analyse of variance*) afin de normaliser les données sans les modifier *a priori*. La seule transformation réalisée avant cette normalisation est le passage au logarithme. D'autres modèles ont été également développés pour les puces à oligonucléotides [113], [100].

Afin d'observer l'effet différentiel des fluorochromes sur les niveaux d'expression mesurés, il suffit de confronter les résultats d'un même échantillon marqué par les deux fluorochromes. Les mesures observées montrent que le fluorochrome affecte les mesures selon les intensités d'expression. Ce biais résulte de différents facteurs dont les propriétés physiques des marqueurs (sensibilité à la lumière et la chaleur, temps relatif de demi-vie), l'efficacité de l'incorporation de ces marqueurs, les différences entre les techniques d'hybridation et les paramètres du scanner [104]. Même si ce biais systématique a une influence relativement faible, il peut mener à de mauvaises interprétations pour l'étude de subtiles différences biologiques [104]. Différentes méthodes sont employées afin de corriger cet effet dont la méthode « lowess » (*Locally Weighted regression Scatterplot Smoothing*) ou encore des régressions locales [3]. Dans tous les cas, les hypothèses sous-jacentes sont identiques aux méthodes globales : peu de gènes varient et leurs variations se compensent globalement [111]. Le lissage lowess repose sur l'idée que toute tendance non-linéaire entre les niveaux d'expression de deux échantillons est un artefact à éliminer [82]. Il est nécessaire de définir le pourcentage de points utilisés afin de lisser les données localement. Plus le pourcentage est grand, plus le lissage est important [104].Tout comme la prise en compte du bruit, il faut utiliser ce type de « correction » avec parcimonie de peur d'introduire du bruit supplémentaire. Ainsi, si l'effet non linéaire est faible, le bénéfice de lisser peut ne pas valoir son coût en terme de bruit rajouté par le lissage [82]. L'idéal non atteint serait de distinguer la non-linéarité artificielle des effets biologiques [82, 100].

Différentes comparaisons entre des méthodes de normalisation ont été effectuées. Il est à noter qu'il s'agit souvent de montrer que la méthode décrite dans l'article est aussi bonne ou surclasse les méthodes couramment utilisées. La comparaison des méthodes lowess, de normalisation globale et celle développée par Zhao *et al.* [112] sur des données simulées montre que, lorsque la majorité des gènes ne sont pas différentiellement exprimés ou que les nombres de gènes sur et sous exprimés sont égaux, la normalisation globale est la meilleure. En revanche, lorsqu'il y a beaucoup de gènes différentiellement exprimés et/ou de façon non symétrique, leur méthode est plus efficace.

Il faut néanmoins prendre ces conclusions avec prudence, puisqu'elles dépendent fortement du modèle de simulation des données ainsi que du critère de comparaison.

Normalisation afin d'adapter la distribution des données à l'analyse

Certaines analyses statistiques classiques reposent sur des hypothèses de distribution de données. Par ailleurs, une distribution fortement asymétrique des données de puces [114] avec un petit nombre de valeurs fortes et beaucoup de valeurs faibles peut poser des problèmes pour l'analyse ultérieure.

Une première étape, souvent négligée avant toute normalisation, est l'analyse de la distribution initiale des données. Hoyle *et al.* [115] ont analysé une variété de données de transcriptome de différents organismes, obtenus à partir de différentes plate-formes et supports. Ils ont montré que les données de transcriptome obtenues suivent toutes une même distribution mixte, proche de la distribution log-normale pour la plupart des mesures ; les extrémités (queues) suivent cependant une distribution du type puissance. La variance de la distribution dépend de la taille du génome étudié. Ces observations sont valables pour toutes les puces globales, c'est à dire pour les puces qui ne présentent pas de biais de sélection des gènes comme les puces dédiées.

Comme la distribution des données est un mélange de deux distributions, il n'existera pas de transformation unique satisfaisante puisque le modèle change selon la valeur des données. Cependant, comme la majorité des données suit une loi log-normale, la normalisation généralement

utilisée est le passage à un logarithme. Cette transformation limite les effets de valeurs extrêmes sur la suite de l'analyse. La distribution obtenue est alors plus proche d'une distribution gaussienne.

D'autres transformations sont possibles, comme l'utilisation de la racine-carrée. Elle ne correspond pas à la distribution des données observées. Sapir et Churchill [116] ont comparé la distribution résultante de la transformation en logarithme et en racine-carrée. Leur conclusion favorise la transformation en logarithme. De plus, cette dernière transformation a l'avantage d'avoir une interprétation biologique. Ainsi, certains phénomènes biologiques ont des effets multiplicatifs sur les niveaux d'expression. Le passage au logarithme permet un passage à des effets additifs qui sont plus facilement modélisés lors des analyses statistiques ultérieures [81]. Cependant, le logarithme pose un problème pour les valeurs négatives obtenues suite à la prise en compte du bruit [73, 81].

Enfin, il est nécessaire de préciser que la recherche de gènes ayant le même profil d'expression peut nécessiter une étape de normalisation particulière suivant la distance utilisée. Si la distance est euclidienne, le niveau d'expression des gènes influencera beaucoup plus les résultats que le profil d'expression. Deux gènes avec de forts niveaux d'expression mais avec des profils différents se retrouveraient, sans normalisation, plus proches que deux gènes exprimés à des niveaux différents mais aux profils identiques.

Quackenbush [108] fournit une information plus détaillée sur les différents types de normalisation.

L'analyse

Les différents types d'analyse seront traités dans le paragraphe suivant. Dans ce paragraphe, seront seulement traités les problèmes inhérents aux données utilisées.

Les valeurs faibles ou bornées

Le problème de fiabilité des valeurs faibles existe pour tous les types de puce à ADN. Les appareils de lecture sont nettement moins précis pour les valeurs faibles [99]. Du coup, les biologistes n'ont que peu de confiance dans ces petites valeurs et les traitent donc, généralement de manière différente.

Elles peuvent être soit purement supprimées soit remplacées par une valeur seuil. Ainsi Tschentscher *et al.* [117] ont remplacé les niveaux d'expression mesurés inférieurs à 50 par la valeur 50 dans des données issues de puces Affymetrix. Cette pratique est également relativement courante dans les puces à lame de verre avec double fluorométrie.

De nombreux jeux de données comprennent, en outre, des valeurs seuils minimales artificielles ou non, des valeurs seuils maximales liées à la saturation du scanner. Lorsqu'un fort pourcentage de gènes présente des intensités seuils, toutes les analyses ultérieures se trouvent biaisées par cette distribution des données. Il arrive parfois qu'il faille réattribuer, à ces valeurs bornées, des valeurs aléatoires afin de pouvoir procéder à des analyses relativement robustes [118].

Les valeurs manquantes

Comme nous venons de le voir, un jeu de données peut comprendre des valeurs manquantes issues soit de problèmes sur la lame (poussières) soit, dans la majorité des cas, de méthodes d'acquisition d'images et de normalisation. La fréquence des valeurs manquantes n'est pas négligeable. Ainsi dans les données de Gash *et al.* [61] 39% des gènes ont au moins une mesure d'expression manquante. On peut atteindre 72,5% pour les données de Garber *et al.* [64] ou 73,5% pour Bohen *et al.* [119].

Or, la plupart des méthodes d'analyse des données requièrent des jeux complets [120]. Afin d'obtenir ce jeu, deux voies sont possibles : soit enlever tous les gènes présentant au moins une valeur manquante, soit réattribuer des valeurs aux données manquantes. Au vu des proportions de valeurs manquantes citées ci-dessus, il est préférable d'utiliser la deuxième voie. L'algorithme de clustering d'Eisen *et al.* [53] présente l'alternative de ne pas tenir compte de la condition expérimentale qui

présente une valeur manquante lors du calcul de distance entre l'expression de deux gènes. Toutefois, cela revient à considérer (pour une distance euclidienne) que l'expression de ces deux gènes est identique dans ces conditions. La distance entre des gènes avec des valeurs manquantes tend donc à être plus faible que les distances entre des gènes qui présentent le jeu de mesures complet [120]. Dans les cas les plus extrêmes, certains clusters obtenus correspondent aux gènes dont les mesures d'expression sont incomplètes.

Généralement, les chercheurs remplacent les valeurs soit par le seuil qu'ils avaient défini au préalable, soit par zéro ou plus rarement, par la moyenne des mesures pour le gène [121, 122]. Ces méthodes ne sont pas optimales puisque aucune ne prend en compte les corrélations existantes avec les autres gènes et que pire, les deux premières incorporent un nombre de valeurs identiques important ce qui biaise toute analyse future. Chiappetta *et al.* [118] proposent de réassigner ces valeurs par des valeurs faibles avec un bruit aléatoire.

Cependant, la réattribution des valeurs manquantes n'est pas une nouveauté propre aux données de puces à ADN. Ainsi, des méthodes classiques comme la réattribution des valeurs par régression [123], par décomposition en valeurs propres (SVD ou singular value decomposition) [124] ou par maximum de vraisemblance sont également utilisées. La plupart de ces méthodes calculent, à partir du plus grand jeu de données complet disponible, des estimateurs des valeurs manquantes. Pour la SVD, les estimations sont fondées sur les valeurs propres ; pour l'algorithme des K voisins les plus proches (K nearest neigbour ou KNN) la valeur est complétée par la moyenne d'expression pour la condition manquante des K gènes les plus proches du gène à compléter (distance euclidienne sur les valeurs d'expression présentes). Beaucoup de méthodes tendent à améliorer le KNN simple : notamment en changeant la distance utilisée, en faisant une moyenne pondérée par la distance des voisins [121], en utilisant les valeurs prédites au fur et à mesure (KNN séquentiel ou SKNN) [125].

La plupart de ces méthodes nécessitent la définition de paramètres comme le nombre de valeurs propres prises en compte pour la SVD ou encore le nombre de voisins utilisés. Plusieurs études ont tenté de définir des gammes de paramètres adéquats. Pour le KNN entre 10 et 20 voisins semblent donner les meilleurs résultats [120, 121, 125]. Toutefois, cette valeur dépend du jeu de données et

notamment du degré de similarité entre les profils d'expression ainsi que du pourcentage de valeurs manquantes. Pour la SVD, les meilleurs résultats obtenus se situent avec la prise en compte d'environ 20% des valeurs propres [121].

Plusieurs comparaisons de l'efficacité de ces méthodes ont été effectuées. La plupart reposent sur un jeu complet de données existantes, auquel un certain pourcentage de valeurs est supprimé aléatoirement. Plus une méthode procure des estimations proche des valeurs réelles, plus elle est caractérisée comme performante (généralement fondée sur le *RMSE root mean square error*). Les méthodes par maximum de vraisemblance présentent des performances bien moindres que le KNN et le SKNN [125]. Pour un pourcentage de valeurs manquantes situé entre 1 et 20%, Troyanskaya *et al.* [121] ont montré que le KNN avec la moyenne pondérée est plus efficace que le SVD et bien plus efficace que la méthode d'attribution de la moyenne du gène. La SVD est plus sensible au type de jeu de données utilisé que le KNN. Pour des jeux de données où plus de 30¨% des gènes présentent au moins une valeur manquante, le SKNN se révèle meilleur que le KNN [125]. Pour un pourcentage de valeurs manquantes égal à 5% l'ACP bayésienne est plus performante que le KNN ou le SVD [120].

La question est de savoir si ces résultats sont extrapolables sur un jeu de données réelles. Un premier point est de définir la taille du jeu de données minimale afin d'avoir des estimations correctes des valeurs manquantes. Le KNN donne des résultats corrects pour un nombre de conditions expérimentales supérieur à 6. En dessous, les estimations sont moins fiables voire mauvaises en dessous de 4 [121].

Comme nous l'avons précisé plus haut, la plupart des données manquantes ne sont pas dues à des poussières sur la puce mais plutôt à un manque de confiance dans les valeurs faibles mesurées. Or, la plupart de ces comparaisons font l'hypothèse de la répartition aléatoire des valeurs manquantes, indépendantes du niveau d'expression du gène. L'efficacité des estimations des valeurs manquantes devrait être moindre dans les conditions réelles [120].

Aussi, la meilleure solution est encore de limiter le nombre de données manquantes, notamment en réduisant ou en éliminant les seuils de détection définis *a priori*. Ainsi, pour toute analyse, le traitement

des données doit être réduit au minimum avant analyse : par exemple, on ne devrait pas enlever les bruit de fond des valeurs ou utiliser des ratios de valeurs car cela ne peut être réalisé sans introduire de biais spécifiques. Similairement, l'introduction de normalisation fondée sur des hypothèses biologiques (moyenne sur des gènes de ménage) introduit un biais systématique [52].

Enfin, dans les expériences d'étude du profil d'expression de l'ensemble des gènes, les variations d'intérêt sont souvent cachées par le bruit [52]. L'interprétation des expériences est gênée par la précision des mesures physiques, de telle manière que l'on néglige souvent l'importance de fluctuations inhérentes aux expériences biologiques. La normalisation des données ne doit pas prendre le pas sur une analyse approfondie qui tient compte des différents facteurs biologiques interférant.

Précisions sur les données de puces

Les intensités des spots ne peuvent pas être prises comme une estimation précise du niveau d'expression mais plutôt comme une mesure relative à l'ensemble des mesures de niveaux d'expression dans l'échantillon étudié [96].

Les données de transcriptome présentent peu de mesures, souvent quelques dizaines de conditions expérimentales, comparées au grand nombre de variables étudiées, les niveaux d'expression de milliers ou de dizaines de milliers de gènes [3]. Par ailleurs, les réplications sont en nombre limité ce qui rend difficile l'évaluation des erreurs et donc, un nombre de faux négatifs et positifs non négligeable [3].

Les niveaux d'expression des gènes ne sont pas indépendants les uns des autres puisqu'ils peuvent participer à un même réseau de régulation ou à une même voie métabolique. A cela se rajoutent des mesures relativement peu reproductibles en raison des différentes sources de variations décrites précédemment.

La confrontation à des données externes

En 1997, une étude sur les effets du stress lié à la température (chaud ou froid) [2] montre que si une minorité de gènes identifiés comme différentiellement exprimés correspondent aux gènes attendus, la majorité ne devrait pas être *a priori* différentiellement exprimée. Afin de savoir si les différences d'expression mesurées correspondent réellement à un changement physiologique, il est nécessaire de prendre en compte des données externes (autres techniques de mesure ou connaissances préalables).

Les résultats obtenus doivent être confrontés à d'autres sources de données avant de se lancer dans des recherches futures consommatrices de temps et d'argent. En effet, les effets observés par l'analyse des puces à ADN ne correspondent pas forcément à l'effet direct escompté. Par exemple, dans une étude de l'influence de mutants de la DAM méthylase sur l'expression des gènes, on s'attendrait à voir des gènes qui présentent des clusters de GATC à l'intérieur de leur séquence ou dans leur région promotrice/régulatrice [126]. Les résultats obtenus ne présentent aucune corrélation entre les gènes différentiellement exprimés et la présence ou non de GATC. L'expérience de transcriptome ne permet donc pas de conclure. Les gènes identifiés peuvent être différentiellement exprimés à cause de l'observation d'effets indirects comme l'activation de voies métaboliques annexes. Ils peuvent également être identifiés comme différentiellement exprimés en raison de changements de conditions expérimentales incontrôlés. Parfois, des phénomènes transitoires au début de l'expérience produisent une empreinte sur la culture pour toute l'expérience et se reflètent dans la mesure des niveaux d'expression [52].

Cette étape, fortement négligée au début de l'engouement pour les puces à ADN, est désormais obligatoire avant toute publication de résultats. Kothapalli *et al.* [84] ont montré que les résultats des puces étaient fortement dépendants de la technique utilisée. Parmi dix-sept gènes supposés différentiellement exprimés, seulement huit (47%) ont été confirmés par Northern blot. Aussi, il s'avère souvent nécessaire, avant de publier ou de se lancer dans une analyse plus précise des gènes détectés, de confronter les résultats des puces avec d'autres sources de données. Une des

possibilités consiste à vérifier la sur ou sous expression du gène d'intérêt à l'aide de mesures par Northern blot ou RT-PCR quantitative [96].

Par ailleurs, les conditions expérimentales, jamais totalement maîtrisées, peuvent révéler le déclenchement de voies métaboliques indépendantes de l'objet de l'étude. La confrontation des données avec les connaissances existantes et notamment les bases de données fonctionnelles permet de valider certains résultats. Des études couplent également analyse du transcriptome et du métabolome chez *Arabidopsis thaliana* [127].

1.3 L'analyse des puces

Cette partie présentera de manière non exhaustive les différents types de méthodes d'analyse de données. Tout comme les méthodes de normalisation, de nombreuses publications portent sur l'analyse des puces avec la création d'algorithmes spécifiques à ces données. A cela se rajoutent les méthodes de statistiques issues d'autres domaines, désormais appliquées au transcriptome. Vu le nombre important de publications à ce sujet, nous aborderons certaines grandes classes de méthodes. Pour une vue plus exhaustive, on peut se référer au site de Li (http://www.nslij-genetics.org/microarray/) qui recense environ 1300 publications sur ce sujet depuis 1993.

La première remarque à faire sur l'analyse des puces est qu'il n'existe, à ce jour, aucun consensus sur la méthode à utiliser [3]. En effet, aucune méthode d'analyse ne permet d'aborder l'ensemble de l'histoire présente dans les données de puces mais uniquement un chapitre de celle-ci [128]. Il est donc nécessaire d'utiliser différentes méthodes plutôt qu'une seule si l'on souhaite tirer le maximum d'information des données disponibles [86].

L'étape de l'analyse est une étape appréhendée par les biologistes, car elle est considérée comme la plus décourageante [3]. Souvent, la crainte de ne pas maîtriser les méthodes statistiques entraîne l'utilisation de méthodes d'abord plus facile comme l'utilisation du ratio et la définition d'un seuil au-dessus duquel un gène est défini comme différentiellement exprimé. C'est d'ailleurs la méthode la plus utilisée dans le monde du transcriptome. Le recours à des méthodes statistiques est cependant

conseillé. Pour cela, la collaboration avec des experts de l'analyse de données est parfois nécessaire mais avant tout, rien ne peut remplacer une formation en analyse afin de mettre correctement en place les expériences [128].

Principes de l'analyse du transcriptome

Les puces à ADN servent généralement à étudier une voie métabolique particulière et à identifier les gènes impliqués dans un phénomène. Or, toute perturbation sur un gène particulier ou sur un composant du signal se propage rapidement à travers le réseau métabolique [129]. Il est en pratique impossible de mettre en œuvre une expérience afin d'observer comment un changement à un nœud du réseau métabolique affectera un autre nœud. En effet, les interconnexions causeront des changements globaux du réseau. Le problème est d'utiliser ces perturbations globales pour retrouver les interactions entre des nœuds individuels.

Le but de l'analyse du transcriptome n'est pas forcément de détecter les gènes avec de grandes variations d'expression, facilement repérables sans statistique. Il faut également rechercher les gènes qui ont des petites variations **reproductibles** [73]. La reproductibilité des résultats est nécessaire avant de se lancer dans des expériences d'approfondissement des conclusions sous peine de perdre du temps et de l'argent dans des efforts non fondés. Dans ce cas, les scientifiques ont besoin d'un plan d'expérience et de méthodes d'analyse qui ne sont pas fondées uniquement sur l'étude relative des mesures mais qui permettent également une estimation des erreurs.

Globalement, l'analyse de puces permet de détecter un grand nombre de gènes comme différentiellement exprimés. Plusieurs articles détectent qu'environ 10% des gènes se trouvent différentiellement exprimés dans les puces, par exemple lors de l'étude du rythme circadien d'*Arabidopsis thaliana* [130] ou de l'influence de mutants GATC chez *E. coli* [126]. Il semble que ce pourcentage de gènes différentiellement exprimés soit une propriété généralement observée dans les études du transcriptome.

L'analyse des puces à ADN comporte deux volets : la recherche de gènes différentiellement exprimés et la recherche de profils d'expression communs. Chaque volet possède des méthodes d'analyse spécifiques.

Les ratios et le principe de seuillage

Dans le reste de ce manuscrit, nous ne traiterons pas des données fournies sous la forme de ratio. Beaucoup d'articles présentent des résultats issus de cette méthode non statistique mais intuitive. Il s'agit de l'utilisation simple du logarithme du ratio entre les mesures de deux conditions expérimentales. Au départ, le ratio a été utilisé pour le marquage par fluorescence avec un échantillon hybridé en vert et l'autre en rouge. Le ratio est désormais appliqué à des marquages différents et même des échantillons hybridés sur des puces différentes. L'idée est que si un gène est différentiellement exprimé entre les deux conditions, son ratio devrait s'écarter fortement de 1. La définition d'un seuil défini arbitrairement, généralement de 3 ou 1/3, permet de définir respectivement les gènes sur/ sous -exprimés dans une condition expérimentale [131]. Elle présente l'avantage d'être intuitive avec une interprétation biologique immédiate : le gène s'exprime x fois plus dans cette condition que dans l'autre. Cependant, cette interprétation est erronée puisque la valeur mesurée comprend une valeur relative de l'intensité d'expression mais aussi des biais liés au protocole expérimental [81]. Les ratios requièrent donc des étapes de normalisation. Les ratios obtenus dépendent alors de la méthode utilisée.

Si elle est aisément compréhensible, elle ne permet pas d'appréhender la complexité des données obtenues. Vu les biais décrits dans le chapitre précédent, les résultats de comparaison d'uniquement deux conditions expérimentales sont nettement insuffisants pour conclure à l'implication de gènes dans le phénomène étudié. Ainsi, si plus de deux conditions expérimentales sont comparées, la méthode ratio seuillage ne suffit plus. Cette méthode ne permet pas d'utiliser l'ensemble des données disponibles et notamment l'utilisation au mieux des éventuels réplicats. Dans le cas où il y aurait réplication, le ratio correspond généralement au ratio des moyennes des mesures.

Par ailleurs, la valeur du ratio est fortement influencée par l'intensité globale de l'expression du gène. On accorde moins de confiance pour de faibles valeurs pour lesquelles le ratio varie plus rapidement [132]. *A contrario,* pour les niveaux d'expression les plus hauts, de petits changements (petits ratios) peuvent être réels mais ils seront rejetés par le seuil défini [133]. Comme une très grande majorité des gènes s'expriment à des niveaux faibles, les ratios d'expression sont aléatoires pour une grande partie de ces gènes. Comme le ratio est moins fiable pour les faibles valeurs, certaines analyses comprennent la suppression des valeurs faibles jugées non pertinentes [108]. On peut ainsi trouver des articles où l'on supprime les valeurs faibles pour ensuite les estimer.

Les ratios sont également utilisés comme mesure de base dans beaucoup de méthodes statistiques appliquées au transcriptome. Si la distribution des intensités d'expression dans une condition peut se rapprocher d'une loi normale, cela n'est pas le cas avec les ratios dont la distribution est plus difficilement modélisable.

Cette utilisation des ratios est donc fortement sujette à caution. Dans la suite de ce manuscrit, les données étudiées sont des niveaux d'expression mesurés pour une condition expérimentale.

Détection de gènes différentiellement exprimés

Généralement, on cherche à détecter les gènes différentiellement exprimés selon le facteur d'intérêt (2 ou plus modalités). Parfois, le plan d'expérience comprend, en plus du facteur d'intérêt, d'autres facteurs biologiques (ex. populations) ou expérimentaux (ex. protocole d'extraction). Dans ce cas, le but est de rechercher des gènes dont le niveau d'expression varie selon le facteur d'intérêt indépendamment des autres facteurs pris en compte.

Il existe de nombreuses méthodes qui servent à l'identification de gènes différentiellement exprimés. L'analyse repose sur les valeurs d'expression d'un gène dans les différentes conditions. Un critère de sélection, spécifique à la méthode employée, permet de déterminer si le gène est différentiellement exprimé ou si la valeur mesurée peut être obtenue par hasard. Les gènes sont étudiés indépendamment les uns des autres, ce qui entraîne un grand nombre de tests par analyse. Deux

problématiques se posent : d'une part, déterminer la distribution des valeurs du critère si les échantillons sont répartis aléatoirement et, d'autre part, prendre en compte le fait que de nombreux tests multiples sont réalisés.

Problématique de la distribution aléatoire des valeurs du critère étudié

Afin de déterminer si un gène est différentiellement exprimé, la valeur du critère mesuré est comparée à la distribution des valeurs du critère lorsqu'il n'y a pas de différence d'expression. Avant l'utilisation massive des calculs *via* ordinateurs, cette distribution était approchée par des lois classiques comme la loi gaussienne ou la loi de Student. Certaines analyses reposent encore sur ce principe d'une loi *a priori* suivie par le critère de sélection.

Cependant, les données de transcriptome s'éloignent généralement des hypothèses formulées (notamment des hypothèses de normalité des résidus). Beaucoup de méthodes d'analyse adaptées au transcriptome simulent la distribution des valeurs du critère de sélection *via* un rééchantillonage aléatoire des données existantes. Les méthodes diffèrent selon les groupes de gènes utilisés pour l'estimation de la distribution. Certains permutent aléatoirement l'ensemble des données tandis que d'autres effectuent des permutations pondérées par la possibilité que ces gènes soient des *outliers* [45].

Problématique des tests multiples

L'analyse du transcriptome porte sur l'étude des niveaux d'expression de milliers de gènes simultanément. Pour chacun de ces gènes, on étudie la probabilité qu'il soit différentiellement exprimé. On fait appel aux statistiques pour répondre à la question : les différences d'expression observées sont-elles bien réelles ? La réponse est indirecte : les statistiques donnent la probabilité qu'il s'agisse d'un faux-positif. Un faux-positif correspond au cas d'un gène où le critère de sélection observé dépasse, par hasard, un seuil fixé à l'avance.

Comme une expérience de transcriptome porte sur des milliers de gènes simultanément, l'analyse statistique est utilisée pour évaluer le pourcentage probable de faux-positifs au-delà d'un seuil donné : 40 gènes dépassent le seuil 1 % par hasard si l'expérience porte sur 4000 gènes alors que c'est le cas de 400 si elle porte sur 40 000 gènes.

L'estimation du nombre de faux-positifs n'est qu'une première étape dans le raisonnement. En effet, on trouve, au-delà du seuil, des faux-positifs et des gènes pour lesquels la différence observée est bien réelle (on la retrouverait dans une autre expérience). L'information clé est la proportion de faux positifs dans l'ensemble des gènes identifiés comme différentiellement exprimés, FDR ou *False discovery rate* [134] car elle mesure le risque de se lancer dans une fausse piste si on décide de travailler sur un gène pris dans cet ensemble. Habituellement le seuil est choisi afin d'avoir moins de 5 % de faux-positifs dans le lot de gènes sélectionnés. Par exemple, prenons une expérience portant sur 4000 gènes et dont 80 gènes sont au-delà du seuil 0,1 %. Comme il y a en moyenne 4 faux positifs au-delà du seuil 0,1 % ($4000 \times 0{,}001$), le pourcentage de faux positifs est de $4/80$, soit 5% des gènes sélectionnés.

Nous avons insisté sur la proportion de faux positifs présents dans la sélection. Bien sûr, idéalement, on voudrait minimiser à la fois les faux-positifs et les faux-négatifs [135], ce qui est irréalisable. Cependant, le plus important est de se concentrer sur les faux-positifs. En effet, ils conduisent potentiellement à une perte d'argent et d'efforts afin de vérifier des relations inexistantes alors que les faux-négatifs représentent des opportunités manquées.

Description succincte de quelques méthodes

SAM (significance analysis of microarrays) [133] identifie des changements d'expression statistiquement significatifs de manière similaire à un test-t pour chaque gène. L'estimation de la distribution de l'estimateur est générée grâce à des permutations aléatoires des données.

D'autres méthodes comme l'ANOVA sont employées de manière classique. Toutefois, comme la distribution de l'erreur obtenue s'écarte généralement fortement de la normalité [73], les probabilités peuvent également être estimées à l'aide de permutations des données.

Détection de profils d'expression semblables

Les méthodes de classification ou de *clustering* visent à identifier des groupes de gènes. L'hypothèse sous jacente est que des gènes co-régulés ou impliqués dans la même voie métabolique doivent avoir leurs niveaux d'expression corrélés.

Problématique du manque de mesures

La plupart des méthodes de classification renvoie systématiquement des groupes de gènes, quelles que soient les données fournies indépendamment de la pertinence biologique [135]. Or, les données de puces sont relativement bruitées à cause des erreurs de mesure et des variations techniques. Toute classification trouvera des profils d'expression dans le bruit autant que dans le signal.

La pertinence d'une règle établie à partir des mesures obtenues dépend du rapport nombre d'échantillons par rapport au nombre de caractères étudiés (ici gène). Les résultats sont généralement estimés robustes pour un rapport d'au moins 5-10 (en fonction des données et de la complexité du classificateur) [136]. Dans le cas des puces, le rapport est typiquement situé en 1/20 (cas rare où il y a beaucoup de conditions expérimentales) et 1/500. Les classifications obtenues ne seront sans doute pas robustes.

Ce problème est d'autant plus important que les résultats obtenus sont la base de diagnostics ou d'explorations plus poussées. Des comparaisons de méthodes de classification ont montré une caractéristique intuitive de la classification : plus le nombre de conditions expérimentales augmente plus les groupes sont significatifs et présentent des gènes co-régulés [135]. A partir de cinquante conditions expérimentales, les groupes sont relativement fiables. Il faut donc prendre garde aux

conclusions réalisées à partir de groupes de gènes obtenus avec peu de données. Pour l'identification de gènes de diagnostic, il faut garder à l'esprit que, lorsqu'il y a un nombre réduit d'échantillons, il est facile d'identifier des classificateurs robustes qui donnent de bons résultats à la fois sur le jeu de données initial et celui de validation. Mais les résultats issus de ces jeux sont illusoires et les conclusions sont suspectes voire fausses [136]. Les jeux de données doivent être suffisamment grands pour être représentatifs de la distribution de la population à étudier par la suite.

Afin d'obtenir une classification pertinente, il est donc nécessaire d'obtenir le plus grand nombre de mesures possibles avec le problème de coût que cela implique et généralement de diminuer le nombre de gènes à classer. La solution conventionnelle est de réduire l'espace des variables/caractères en éliminant les informations redondantes ou les éléments bruités. Dans notre cas, il est possible d'éliminer les gènes dont les niveaux d'expression ne varient pas ou très peu lors des conditions expérimentales [118]. Ce préalable est cependant souvent négligé.

Une des idées proches de ce problème est d'identifier jusqu'à quelle profondeur un arbre de classification est fiable. Cette profondeur de fiabilité dépend, contrairement au problème précédent, de la méthode de classification utilisée. Ainsi, la classification hiérarchique donne des résultats fiables pour des groupes de moins de dix gènes [122]. D'autres méthodes permettent d'obtenir des résultats fiables pour des groupes constitués de moins de cinquante gènes, rarement plus.
Certaines études présentent une interprétation des résultats fondée sur une classification globale de l'ensemble des gènes. Les conclusions qui en découlent ne sont généralement pas fiables.

Problématique de la définition de similarité ou distance entre les gènes

Une autre difficulté est de présenter la similarité entre les gènes (qui correspond à un espace d'une dizaine à plusieurs dizaines de dimensions) sur un espace à deux voire trois dimensions. Chaque méthode nécessite forcément des compromis [122] et chaque méthode explorera le nuage des données de manière particulière.

Cette exploration du nuage dépend de la distance entre deux gènes utilisée : distance euclidienne, angle, corrélation linéaire, … D'autres mesures de similarités ont été développées. Elles prennent en compte un phénomène d'apprentissage entre deux jeux de données : des données de transcriptome et des données auxiliaires comme des classes fonctionnelles ou des lieux d'expression des gènes [122]. La distance la plus utilisée est la distance euclidienne. En présence d'une normalisation sur le niveau d'expression de chaque gène, elle correspond à la corrélation linéaire. Dans le cas contraire, elle mesure à la fois la différence de niveau d'expression et le profil d'expression. Aussi, elle conduit généralement au rassemblement des gènes de même niveaux d'expression. Le coefficient de Pearson permet de mettre en évidence une similarité de profils d'expression sans tenir compte des niveaux moyens d'expression [135]. Yeung *et al.* ont comparé le nombre de gènes qui présentent le même facteur de régulation au sein des clusters obtenus. Quelle que soit la méthode utilisée, la corrélation offrait de meilleurs résultats que la distance euclidienne [135] sans normalisation préalable.

Différents types de méthodes de classification

En plus de multiples possibilités de mesures de distance, il est possible de distinguer deux types de classification : la classification supervisée où l'on connaît déjà les conditions discriminatoires (par exemple différents types de cancer) et la classification non supervisée qui présente une démarche plus exploratoire.

Dans la classification supervisée, le but est d'obtenir :

- des groupes robustes (indépendants d'éventuel outliers) tels que les nouveaux échantillons soient classés correctement
- un nombre réduit de gènes discriminants.

Le but final est généralement de définir un jeu de gènes marqueurs d'un type d'échantillon (type de cancer) avec un nombre de gènes limités pour pouvoir réaliser une puce dédiée au diagnostic mais efficace pour le diagnostic. Malheureusement, il est assez difficile d'avoir ces deux critères réunis [136]. Généralement, les gènes marqueurs sont identifiés indépendamment les uns des autres : les M meilleurs sont sélectionnés. Ces M gènes peuvent receler des informations communes et donc inutiles

[136]. Une autre solution est d'identifier un nombre important de jeux de gènes ayant un pouvoir de classification et, ensuite, de compter l'occurrence de chaque gène dans ces jeux et de prendre les plus fréquents.

La classification non supervisée n'a pas d'information *a priori* sur les groupes de gènes à identifier. La classification peut être ascendante ou descendante. Certaines méthodes requièrent, cependant, le nombre de clusters à identifier comme les méthodes de K-means ou les SOM (*self organizing maps*) alors que d'autres produisent des arbres de classification sans définition de nombre de groupes.

Une des méthodes les plus utilisées sans définir le nombre de groupe *a priori* est la classification hiérarchique qui rassemble les gènes les plus proches en un groupe, calcule les niveaux d'expression représentant ce groupe et poursuit ensuite la classification [135]. La méthode la plus utilisée est la méthode graphique de Eisen [53]. Son interface graphique et sa facilité d'utilisation ont encouragé son utilisation. Le graphique comprend, d'un côté l'arbre de classification des gènes obtenu par classification hiérarchique avec comme distance la corrélation linéaire et, de l'autre le graphique des résultats obtenus sous forme de couleur rouge, vert ou jaune selon les conditions expérimentales (Figure 9). Les groupes sont définis en coupant arbitrairement à un niveau de l'arbre.

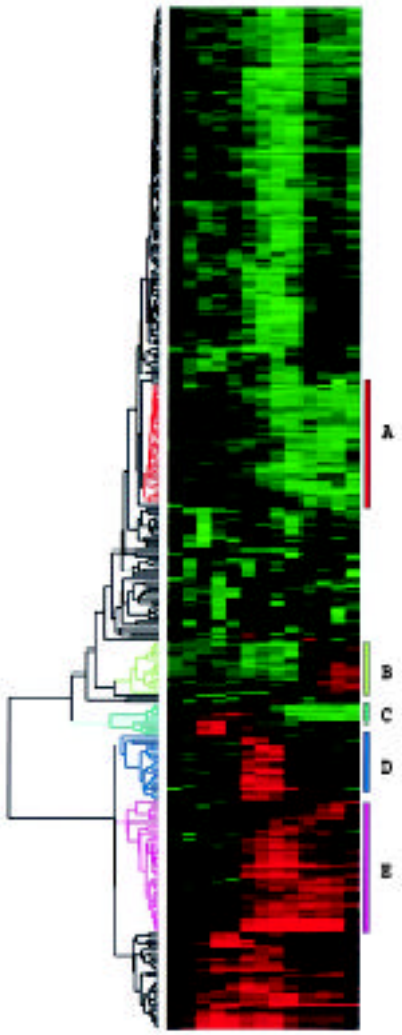


Figure 9 : Exemple de graphique de Eisen (tiré de [53])

Eisen *et al*. [53] reconnaissent, à la fin de leur article, que la méthode de classification utilisée n'est pas forcément la plus adéquate. Depuis, de nombreux articles ont été publiés présentant des méthodes généralement importées d'autres domaines scientifiques. Ces méthodes vont de la plus simple comme la classification hiérarchique à la plus sophistiquée. Comme l'ont fait remarquer Somorjai *et al.* [136], la maxime « le plus simple est le mieux » a été la plupart du temps ignorée. En général, aucun effort n'a été réalisé afin de choisir le classificateur le plus approprié selon le type de jeu de données. Le choix s'effectue soit en prenant l'exemple d'un autre article, l'expérience et les préférences personnelles ou la disponibilité du logiciel. Et pourtant la complexité du classificateur et la taille de l'échantillon doivent être corrélées.

Problèmes de méthodes de classification et méthodes exploratoires

Un des problèmes des méthodes de classification est que chaque gène sélectionné est présent dans la classification finale (même si son profil d'expression est relativement éloigné des autres). Par ailleurs, il n'est présent que dans un seul groupe de gènes [137]. Or, biologiquement, un même gène peut participer à plusieurs voies métaboliques ou de régulation définis dans différents clusters. On s'attend à ce que chaque gène soit influencé par plusieurs facteurs de transcription et qu'il puisse influencer différents gènes [118].

D'autres méthodes sont donc utilisées afin d'identifier des groupes de gènes dont les profils d'expression sont proches : ce sont des méthodes généralement utilisées afin de visualiser le nuage des données. Ainsi, l'analyse en composante principale (ACP) [138] ou l'analyse en composantes indépendantes (ACI) ont donné des résultats intéressants dans le cadre du transcriptome [118]. Ces deux méthodes recherchent des axes qui contiennent la plus grande part de l'information contenue dans les données. L'ACP recherche des axes orthogonaux qui représentent les plus grandes variances dans les données et l'ACI des axes statistiquement indépendants sur lesquels les données s'écartent le plus de la loi normale qui caractérise généralement le bruit. Ces axes sont supposés représenter des processus biologiques indépendants. Comme ces méthodes ne sont pas dirigées, les axes ne représentent pas forcément les facteurs d'intérêt. Toutefois, dans notre expérience et dans les différents articles qui comprennent cette méthode, un axe ou un plan représentant les facteurs d'intérêt de l'expérience ont toujours été trouvés. Pour l'ACP, le premier axe représente toujours les niveaux d'expression relatifs de chaque gène [138], soit entre 80 et 90% de la variance des données. Cette proportion est d'autant plus forte que l'expérience porte sur de petites fluctuations des niveaux d'expression avec peu de gènes impliqués. L'ACP est également une méthode dédiée à l'étude de phénomènes temporels.

Ces méthodes peuvent également être utilisées pour réduire l'espace des données. Il suffit pour cela de filtrer les axes qui rassemblent du bruit ou des artefacts et de garder les axes qui contiendraient potentiellement l'information biologique. Il suffit d'appliquer ensuite des méthodes de classification ou, tout simplement, de visualiser les groupes de gènes discriminés par les axes restants.

L’article qui suit complète cette approche bibliographique et détaille notamment cinq méthodes d’analyse utilisées sur les données de transcriptome : l’ACP, l’ACI, le t-test, l’ANOVA et SOM. Il présente également différentes questions souvent posées par les biologistes lors des formations permanentes auxquelles nous avons participé. Enfin, il souligne le fait qu’aucune méthode de normalisation et d’analyse ne peut compenser la nécessité d’un plan expérimental bien établi.

Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Biology and Chemistry
Computational Biology and Chemistry 29 (2005) 319–336
www.elsevier.com/locate/compbiolchem

Comments on selected fundamental aspects of microarray analysis

Alessandra Riva [a,*], Anne-Sophie Carpentier [a], Bruno Torrésani [b], Alain Hénaut [a]

[a] *Laboratoire Génome et Informatique UMR 8116 Tour Evry2, 523 Place des Terrasses, 91034 Evry Cedex, France*
[b] *LATP, CMI, Université de Provence, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France*

Received 26 May 2005; received in revised form 18 August 2005; accepted 18 August 2005

Abstract

Microarrays are becoming a ubiquitous tool of research in life sciences. However, the working principles of microarray-based methodologies are often misunderstood or apparently ignored by the researchers who actually perform and interpret experiments. This in turn seems to lead to a common over-expectation regarding the explanatory and/or knowledge-generating power of microarray analyses.

In this note we intend to explain basic principles of five (5) major groups of analytical techniques used in studies of microarray data and their interpretation: the principal component analysis (PCA), the independent component analysis (ICA), the *t*-test, the analysis of variance (ANOVA), and self organizing maps (SOM). We discuss answers to selected practical questions related to the analysis of microarray data. We also take a closer look at the experimental setup and the rules, which have to be observed in order to exploit microarrays efficiently. Finally, we discuss in detail the scope and limitations of microarray-based methods. We emphasize the fact that no amount of statistical analysis can compensate for (or replace) a well thought through experimental setup. We conclude that microarrays are indeed useful tools in life sciences but by no means should they be expected to generate complete answers to complex biological questions. We argue that even well posed questions, formulated within a microarray-specific terminology, cannot be completely answered with the use of microarray analyses alone.

Keywords: Microarrays; Fundamental tools; FAQ section

1. Introduction

Microarrays have become one of the fundamental tools for biologists and great hopes are placed in their ability to answer all the questions asked by the researchers.

The amount of data created in an experiment is large and the nature of the data quantitative, two features a biologist is not necessarily used to or trained for. For the analysis of the data, the biologist has to choose from a rapidly increasing number of methods proposed in the literature, again, without necessarily having the knowledge and competence to do so. He therefore risks overestimating the power and capacity of the method (to provide him with the answers he is looking for).

This commentary wants to give the fundaments, which will allow the biologist to get out a maximum from microarrays, by understanding their nature and the principles of the statistical methods proposed to him.

For this we first give a brief introduction to the subject of microarrays, their origins, the different types and their application. We then examine the fundamental groups of methods used in the analysis of microarrays. Throughout we provide the reader with a list of papers allowing him to pursue the point further.

The FAQ section, which follows, contains the answers to questions related to the analysis of microarray data, often asked during the course taught by this laboratory (http://www.infobiogen.fr). This is another way to approach the subject and again, a list of publications for the interested reader is provided.

The last section leads us to consider which are the important aspects in the experimental setup, in function of the analysis methods discussed.

* Corresponding author. Tel.: +33 1 60 87 38 63; fax: +33 1 60 87 38 97.
E-mail address: gucki@genopole.cnrs.fr (A. Riva).

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.08.006

2. Fundamentals and basic terminology

2.1. General introduction to microarrays

A microarray consist of a solid support on which a series of DNA segments is arranged and fixed in a regular pattern. These segments are incubated with a labelled nucleic acid sample. When a nucleic acid sequence in the sample is complementary to a DNA segment present on the support, it will bind and hybridize to this, specific segment. This hybridization is recorded and analyzed.

2.1.1. The historical background

As Jordan (2002) points out, DNA arrays were already being used in the seventies, in the form of dot blots and slot blots. Ekins and co-workers developed microspot fluorescent immunoassays in the late eighties and early nineties, proving that the sensitivity of these miniaturized assays was comparable to that of "macroscopic" ones and introducing the concept of microarray (Ekins, 1989; Ekins et al., 1990; Ekins and Chu, 1991). The concept of miniaturization was also applied to DNA arrays, using two different approaches. One was to deposit the DNA (or complementary DNA) on glass plates, leading to the first publication of a gene expression microarray article in 1995 (Schena et al., 1995). The second approach was that of the oligonucleotide array, where the DNA is directly synthesized onto the support (Fodor et al., 1991; Southern et al., 1992).

2.1.2. Today's microarrays

In the following, "probe" denotes the immobilized DNA on the support and "target" the mobile DNA, cDNA or mRNA. Some authors, however, use the terms the other way round. The supports used for microarrays today are glass (microscope) slides (nylon) membranes or silicon chips.

The material fixed on the support ("probe") can be:

a. DNA, representing coding sequences or, more generally, pieces of genomic DNA.
b. Complementary DNA, obtained from the mRNA of specific genes or expressed sequence tags (ESTs); the latter is usually used for organisms not yet completely sequenced.
c. Oligonucleotides; in the case of oligonucleotide arrays the oligos are synthesized directly onto a silicon chip; this process has been pioneered by Affymetrix (see Lipshutz et al. (1999) for a comprehensive review on oligonucleotide arrays).

The mobile "target" can be:

a. DNA.
b. Complementary DNA (cDNA), obtained from mRNA by reverse transcriptase-PCR (RT-PCR).
c. mRNA; this can be used although cDNA is generally preferred.

A hybridization mixture is obtained by labelling the target fluorescently or radioactively. This mixture is then incubated with the prepared microarray and allowed to hybridize with the probe. Finally, the resulting signal intensity, which correlates with the amount of captured probe, is measured, stored in a computer and then analyzed.

Recently, efforts have been made to extend the microarray technology to the field of proteins. The interested reader may refer to the review written by Templin et al. (2002) for a comprehensive introduction to this field. For further information on microarray technology, the reader may refer to recent review articles (Barrett and Kawasaki, 2003; Vrana et al., 2003); he may also refer to a related NCBI web page (http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html).

2.2. Applications

Microarrays can be used for a variety of purposes, including the detection of mutations, DNA sequencing and the analysis of gene expression. Microarrays allow measuring the expression levels of thousands of genes at the same time and this opens the possibility to identify differentially expressed genes (Callow et al., 2000) and to cluster those genes sharing similar expression patterns (Heyer et al., 1999). They have become a widespread tool for analyzing the relative transcription levels of genes.

The fields microarrays are being used in are numerous and constantly growing, some examples being:

a. clinical medicine (see Joos et al. (2003) for a review on this subject);
b. the study of the cell-cycle (see for example McCune and Donaldson (2003));
c. the study of the circadian rhythm in animals (see for example Stanewsky (2003)) and plants (see for example Davis and Millar (2001));
d. the study of plant metabolism (see for example Hirai et al. (2004)).

They are also being used to elucidate the role of non-coding sequences, for example, the role of some promoter regions, by integrating expression profiles with the information on promoter sequence similarity (Bussemaker et al., 2000; Park et al., 2002). Heterologous hybridization to cDNA microarrays is gaining in popularity and is, for example, used in order to elucidate the molecular basis of complex traits in "non-traditional model systems" (Renn et al., 2004).

As different as these applications may seem, the aim of the experiments is one of the following:

a. To find the genes which indicate a phenomenon (not necessarily at the origin of the phenomenon, but an indicator of it: expression change correlated with the phenomenon).
b. To find the genes which are at the origin of the phenomenon under investigation.

In the first case, the researcher will need to find genes whose expression levels change considerably, few in numbers and that can be preferably used in antibody assays (still

cheaper and faster to set up than microarrays) (Deutsch, 2003); an analysis of the microarray data will generally be sufficient in order to identify the genes.

The work done by Sekowska et al. (2001) and Oshima et al. (2002) are examples for the second case and we will come back to them in the course of the commentary; here, an analysis of the microarray data is not enough to find the genes at the origin of a phenomenon (and only these): it is necessary to combine the results of the microarray analysis with information from other sources, such as the genomic and the purely biological fields (Jarvis et al., 2004; Hirai et al., 2004; Riva et al., 2004). This is something important to bear in mind and will be discussed at various points of the commentary.

3. Data representation and analysis

3.1. The raw data

The microarray data used in the following stem from experiments on the sulphur metabolism of *B. subtilis* (Sekowska et al., 2001) and are freely available at http://195.221.65.10:1234/∼carpenti/. The experiments were carried out using Panorama nylon filters *B. subtilis* gene arrays (Sigma-GenoSys Biotechnologies); each array contains all of *B. subtilis*' genes and one gene is represented by one spot. Each gene spot is represented twice on the array.

The aim of these experiments was to identify the genes differentially expressed when the bacteria are grown with methionine ("met") or methyl-thioribose ("mtr") as sulphur source. The experiments followed a fully crossed factorial design with four factors (sulphur source, day of experiment, amount of RNA used and duplicate of each spot). The data (raw levels of expression) were gathered in an array of 4107 rows (all *B. subtilis* genes) and 16 columns (experimental conditions). The minimum value was 213, the maximum value 13,455, with two thirds of the data having a value below 800. Note that each factor has only two states: all factors are binary (see Fig. 1).

3.2. The data table and some preliminary considerations and manipulations

It is natural to want to represent the data in a graph. We obtain one (and only one!) graph, with N dimensions (corresponding to the N experimental conditions), a cloud in an N-dimensional space. As we are not good at coping with drawings having more than two dimensions (three still works well on a computer screen), we are obliged to take the columns 2 by 2 (i.e. one experimental condition versus another).

Note that when you draw a graph by hand, you will automatically try to maximize the use of the paper: you look at the minimum and maximum values for both variables, and define the scale accordingly. The machine will do the same. In both cases, the data are transformed through a change of variable: 1 cm on the graph corresponds to X units of the original variable (a linear transformation).

3.2.1. Translation

This is an operation which in itself does not pose a problem, as one is interested in the relative position of the points to each other: the aim is to find the points that are far away relative to the main body of the cloud, which means that the reference frame used to look at the cloud does not really have much importance. However, the translation may create complications when it consists in bringing a lot of the values close to zero followed by taking the log of the data, something discussed in the next section.

3.2.2. Normalization

Note that drawing a graph or letting a spreadsheet (like MS-Excel) draw the graph, implicitly presumes that the sum of the signal does not change in function of the experimental conditions; one allows the data to be normalized. By doing this, one has presumed that the total of the signal in each column is the same: total signal of column 1 = total signal of column 2. This is justified when three conditions are fulfilled: firstly, more than 90% of the genes do not care about the experiment, i.e. do not change expression in function of the different experimental conditions; in that case one can indeed presume that the total quantity of cDNA (and therefore of the mRNA) is the same. Secondly, the number of genes analyzed has to be large: this is a way to make sure that the majority of the genes do not change expression in function of the different experimental conditions. Thirdly, the overall intensity change of up- and down-regulated genes is similar. The three conditions are fulfilled in our example, but they would not be in, say, the temporal analysis of mRNA decay. The reader is referred to the work of Stoyanova et al. (2004) for some interesting considerations on this subject, as well as to the work

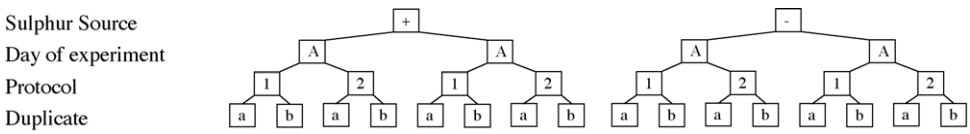


Fig. 1. Experimental design of the transcriptome analysis on *Bacillus subtilis* (Sekowska et al., 2001). The experimental setup follows a fully crossed factorial design. In the case of Sekowska et al. (2001) the quantity of RNA used for the RT-PCR differed between the two protocols. Note that changing the protocol (a different quantity of RNA or labelling with Cy3 rather than Cy5) or having duplicats for each gene on the array are all technical factors which increase the workload without adding any biologically pertinent information. It is preferable to increase the number of states for the biological factors, in the above case an additional sulphur source or an additional experimental day.

of Zhao et al. (2005) who propose a normalization procedure for data not fulfilling the above conditions.

Instead of just looking at the minimal and maximal values in order to best represent the graph, it is advisable to calculate the means and variance for each experimental condition: in the first case the estimates are based on two points only (min and max) per experimental condition, in the second case the estimate is made using all points. If these are numerous, the result is more stable.

3.3. Graphic exploration

3.3.1. Preliminary considerations

As we said, we are forced to take the columns 2 by 2, which means that we will look at *projections of our single cloud on the different planes*.

What are we looking for? Presuming that the three above-mentioned conditions are fulfilled, at least 90% of the genes analyzed will not change expression under the different experimental conditions. This means that on the graph one would see them all lying on one line, if it was not for the noise: the noise is responsible for making those points look more like a cigar which is the wider the more noise there is. The remaining 10% of the genes will change expression; they have an atypical behaviour and will not lie on the line (the cigar) but be apart. These genes that are apart from the main body of the cloud are the ones the biologist is interested in. Note that having the 90% of the point lying on a line is an ideal case, the "cigar" being the reality; so one tries to find that line (which describes 90% of the genes) somehow.

How do we describe those 90% of the data? How do we determine the line? Various options are available:

(a) One can try to draw it by hand.
(b) Calculate the linear regression. This is not such a good idea as there are two lines of regression (x-axis versus y-axis and vice versa) and they are not identical except when all the points lie on the same line.
(c) Use methods that are more sophisticated.

The methods all presume that the cloud follows a Gaussian distribution, or at least a unimodal and symmetrical one. They also need some pre-processing of the data, for two reasons:

(a) The fact that the data often consist of a very large amount of small values and a few, extreme points, something which affects most data analysis techniques strongly (Chiappetta et al., 2004).
(b) Some effects being studied may have a multiplicative behaviour.

To solve the first of these problems, taking the log, the square (or cubic or fifth etc) root or the hyperbolic tangent are all possible and generally accepted methods (see Fig. 2), whilst for the second problem taking the log is preferable (Chiappetta et al., 2004; Hoyle et al., 2002; Thygesen and Zwinderman, 2004; Tusher et al., 2001).

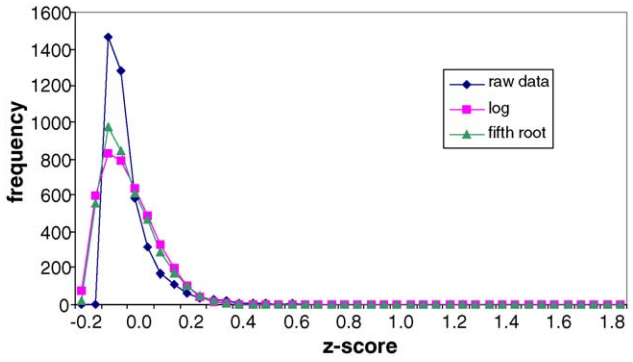


Fig. 2. Effect of different pre-processing methods on the data distribution. The figure shows the effect different pre-processing methods have on the data distribution. Shown are the distributions (◆) of the raw data, (■) after having taken the log and (▲) after having taken the fifth root. As can be seen, either operation brings the distribution closer to a Gaussian one.

As we mentioned in the section above, the reference frame used to look at the cloud does not really matter and making a simple translation does not in itself pose a problem. One does need to be careful, though: making a simple translation is indeed no problem, nor is taking the log. However, when executing both operations, one needs to be cautious: if the translation consists in bringing a lot of the values close to zero, taking the log afterwards will create a distortion in the cloud of points: one has just created a whole package of data with values going towards minus infinity. This means that in trying to take care of the problem of the points at the far right (few points with very large values) by taking the log, the result is worse than the starting point. Note that when executing the two operations in the inverse order (first log, then translation) the problem is not created.

We come back to the graphs, which are just many projections on different planes of ONE cloud. A brief look at the general shape of each cloud projection is worthwhile. If a cloud resembles a fat cigar, a lot of genes have considerably changed expression. If, on the other hand, the cloud resembles a line, the great majority has not changed expression (see Fig. 3 for two examples). The "cigar" may also be bent or twisted. In this case the readings were taken outside the linear range of the machine, an issue discussed in Section 4.1.7. We can be faced with a problem: taking the columns 2 by 2, the number or graphs increases very rapidly when increasing the number of experimental conditions: in our example we have 16 columns which means we need to look at $16 \times 15/2$ i.e. 120 graphs. Evaluating them all in detail becomes a bit tedious.

Thus, we need to find ways to reduce the number of graphs we have to examine. To do this, we need to decide, from which point of view we want to look at the cloud, which has to be translated into a mathematical criterion. This implies that there will be a change (rotation) of the reference frame.

3.3.2. By hand (with a spreadsheet)

With "by hand", we refer to the fact that the calculations are extremely simple. As the calculations have to be repeated

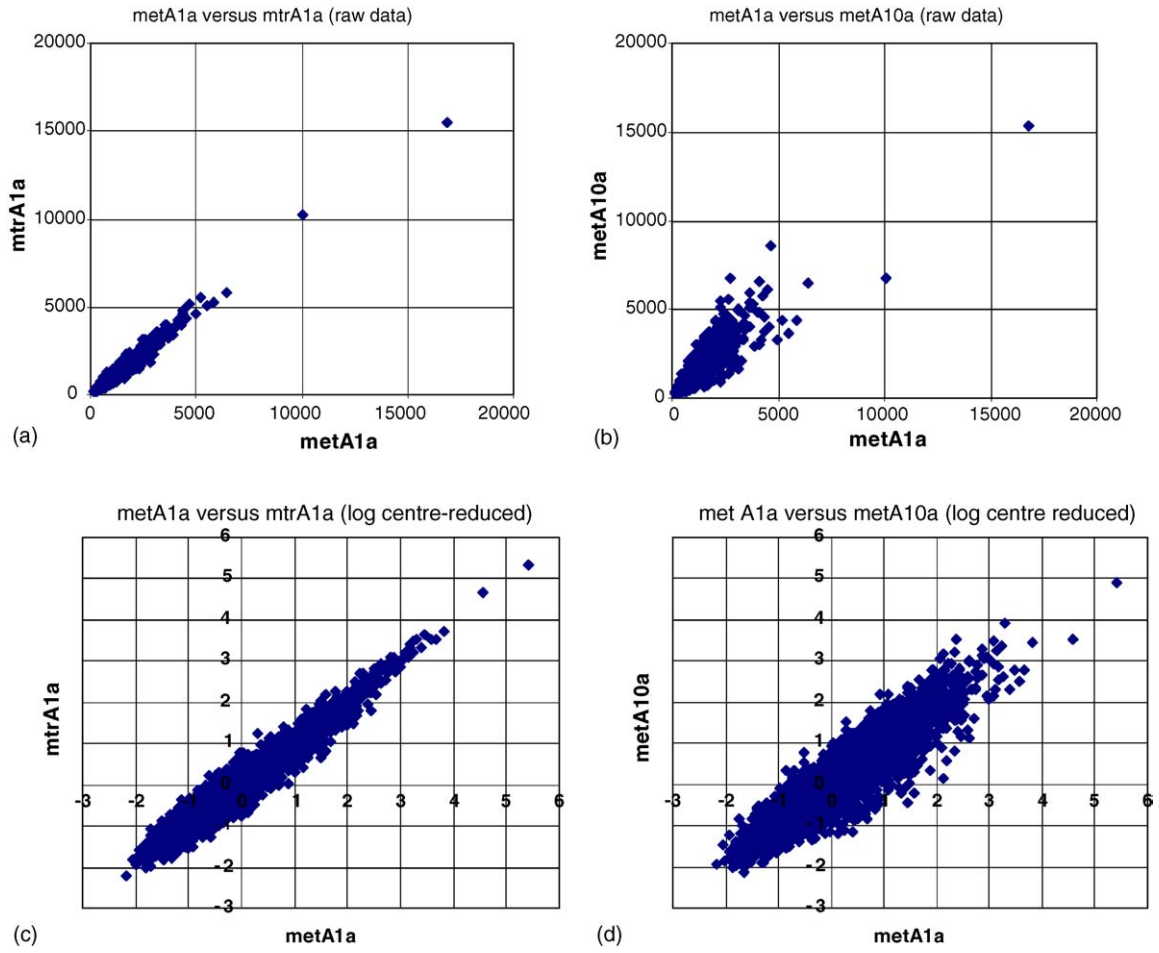


Fig. 3. Projections of the data on different planes. In all four figures, each axis corresponds to an experimental condition: (a and c) metA1a vs. mtrA1a; (b and d) metA1a vs. metA10a (see Fig. 1 for the nomenclature). (a) and (b) show projections of the raw data, in (c) and (d) the data are log centre-reduced. Log centre-reducing the data has brought the few points which are far away from the main body in (a) and (b) closer in (c) and (d). Note how the space is more efficiently used in (c) and (d). The points in the two left-hand pictures form a narrower "cigar", indicating that fewer genes have changed expression than on the right-hand side.

for each gene, though, the number of calculations is such as to make handing the job over to a spreadsheet a practical alternative.

The only reasonable option to reduce the 120 little pictures means concentrating on the expression changes caused by each single factor being studied, in our case four. For this we calculate the mean expression for each gene; this will be the *x*-value. Then, for a given factor like sulphur, we calculate the sum of all met values and subtract from it the sum of all mtr values, which gives us the *y*-value.

This is done for all four factors. Note that we have changed the reference frame; this calculation, which is done instinctively by hand, can be formalized and done via a matrix, called "mixing matrix": it allows to change from the old reference frame to the new one and is shown in Table 1.

We obtain four graphs, one for each factor; we then look for genes that are far away from the main body of the cloud. Fig. 4 shows the graph obtained for the factor sulphur. Executing this operation, each experimental condition is given the same weight and the criterion chosen to look at the cloud is "one factor per graph".

3.3.3. PCA

A more sophisticated approach is the principal component analysis. Pearson first introduced it in 1901. The reader may refer to the work by Stoyanova et al. (2004) for a comprehensive introduction to the subject and to Kendall et al. (1983) for a technical presentation.

Here, the criterion chosen to look at the cloud is to maximize the variances along the axes of the reference frame. There are numerous softwares that do this job and which supply us with the mixing matrix, which in PCA's case is called *eigenvector matrix*, shown in Table 1. This matrix allows us to change from the old reference frame to the new one; it gives us for each of the new axes (in the table: the columns) the coefficient with which we have to multiply each gene's value in a given experimental condition (in the table: the lines) in order to obtain its new coordinates (see legend of Table 1).

Table 1
The mixing matrix calculated by the spreadsheet (MS-Excel) and the *eigenvector matrix* calculated by PCA

Experimental condition	Axis				
	Mean expression	Effect of protocol	Effect of day	Effect of sulphur source	Effect of duplicate
metA1a	0.250	0.250	-0.250	0.250	-0.250
metA1b	0.250	0.250	-0.250	0.250	0.250
metB1a	0.250	0.250	0.250	0.250	-0.250
metB1b	0.250	0.250	0.250	0.250	0.250
metA10a	0.250	-0.250	-0.250	0.250	-0.250
metA10b	0.250	-0.250	-0.250	0.250	0.250
metB10a	0.250	-0.250	0.250	0.250	-0.250
metB10b	0.250	-0.250	0.250	0.250	0.250
mtrA1a	0.250	0.250	-0.250	-0.250	-0.250
mtrA1b	0.250	0.250	-0.250	-0.250	0.250
mtrB1a	0.250	0.250	0.250	-0.250	-0.250
mtrB1b	0.250	0.250	0.250	-0.250	0.250
mtrA10a	0.250	-0.250	-0.250	-0.250	-0.250
mtrA10b	0.250	-0.250	-0.250	-0.250	0.250
mtrB10a	0.250	-0.250	0.250	-0.250	-0.250
mtrB10b	0.250	-0.250	0.250	-0.250	0.250

Experimental condition	Axis															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
eigenvalue	94.64	2.26	1.01	0.56	0.36	0.29	0.26	0.18	0.16	0.05	0.05	0.05	0.04	0.04	0.03	0.02
metA1a	0.250	0.307	-0.164	-0.114	0.336	-0.004	-0.310	0.179	0.349	-0.146	-0.130	0.406	-0.228	0.429	-0.020	-0.017
metA1b	0.249	0.312	-0.166	-0.131	0.366	0.075	0.157	0.252	0.347	0.131	0.078	-0.429	0.262	-0.413	0.001	0.026
metB1a	0.252	0.179	0.115	0.365	0.124	0.306	-0.301	-0.280	-0.158	0.172	0.629	-0.094	-0.019	0.132	0.052	0.028
metB1b	0.251	0.181	0.107	0.369	0.153	0.425	0.211	-0.246	-0.203	-0.170	-0.592	0.072	0.003	-0.140	-0.036	-0.027
metA10a	0.249	-0.256	-0.326	0.211	0.200	-0.429	-0.183	-0.129	-0.186	0.372	-0.309	-0.254	0.193	0.264	0.082	0.028
metA10b	0.249	-0.255	-0.328	0.192	0.260	-0.303	0.315	-0.027	-0.146	-0.388	0.325	0.261	-0.198	-0.276	-0.081	-0.031
metB10a	0.250	-0.254	0.263	-0.232	0.126	0.086	-0.322	0.311	-0.322	-0.041	-0.070	-0.072	-0.146	-0.163	-0.330	0.505
metB10b	0.250	-0.253	0.259	-0.236	0.178	0.180	0.170	0.358	-0.283	0.064	0.067	0.092	0.171	0.180	0.335	-0.508
mtrA1a	0.248	0.334	-0.083	-0.365	-0.224	-0.182	-0.173	-0.256	-0.270	0.108	0.003	0.432	0.363	-0.310	0.093	0.018
mtrA1b	0.248	0.331	-0.087	-0.370	-0.189	-0.093	0.321	-0.159	-0.277	-0.089	0.029	-0.407	-0.399	0.305	-0.086	-0.028
mtrB1a	0.251	0.156	0.220	0.311	-0.360	-0.291	-0.164	0.272	0.049	-0.540	0.002	-0.211	0.315	0.102	0.061	-0.029
mtrB1b	0.252	0.149	0.214	0.322	-0.290	-0.206	0.256	0.332	0.099	0.540	-0.012	0.236	-0.301	-0.101	-0.066	0.029
mtrA10a	0.249	-0.239	-0.346	-0.031	-0.376	0.270	-0.285	0.035	0.151	-0.042	-0.089	-0.114	-0.350	-0.261	0.475	-0.045
mtrA10b	0.249	-0.240	-0.352	-0.030	-0.342	0.369	0.195	0.052	0.150	0.053	0.087	0.108	0.352	0.262	-0.471	0.045
mtrB10a	0.251	-0.226	0.335	-0.131	-0.003	-0.151	-0.196	-0.375	0.341	0.014	-0.054	-0.094	-0.118	-0.174	-0.385	-0.484
mtrB10b	0.251	-0.225	0.327	-0.140	0.041	-0.057	0.310	-0.322	0.358	-0.037	0.036	0.067	0.101	0.164	0.376	0.491

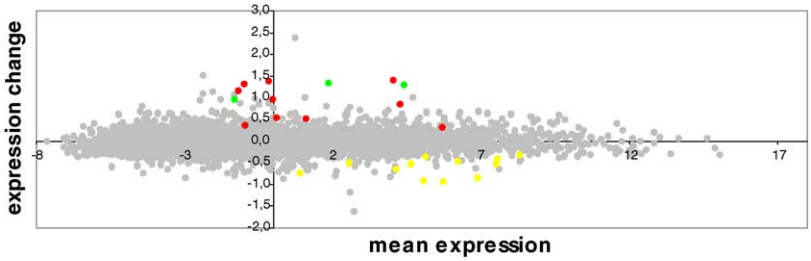


Fig. 4. The expression change in function of the factor sulphur as calculated by a spreadsheet (MS-Excel). The figure shows the genes' expression change in function of the sulphur source against their mean expression. The potentially interesting genes are those away from the main body of the cloud. The highlighted genes are the ones which proved to be of particular interest for the problem investigated by Sekowska et al. (2001); the reader may refer to their work for a detailed discussion. Note that not all of these genes would have been detected using the spreadsheet.

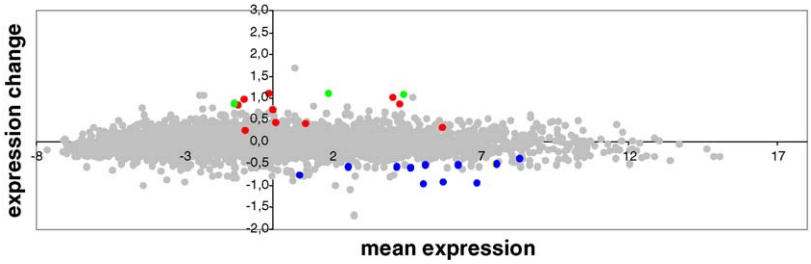


Fig. 5. The data cloud projected on the plane formed by axis 1 against axis 5 (PCA). The figure shows the genes expression change in function of the sulphur source (axis 5) against their mean expression (axis 1), as calculated by PCA. The potentially interesting genes are those away from the main body of the cloud. The highlighted genes are the ones which proved to be of particular interest for the problem investigated by Sekowska et al. (2001); the reader may refer to their work for a detailed discussion. Note that not all of these genes would have been detected using PCA.

The *eigenvector matrix* gives us also another information: the variance or *eigenvalue* for each axis, expressed in percentage. This provides an indication of the cloud's dispersion along the axis (the bigger the value, the more the genes are dispersed along this axis). The fundamental idea is that if the dispersion is great, the image is easier to interpret than if all the points were packed together. If an experimental factor influences the expression of some genes, the factor will contribute to the dispersion of the cloud and may coincide with one of the axes determined by PCA.

The *eigenvector matrix* gives a wealth of information. Looking at our matrix, we see that for the first axis all the 16 coefficients have basically the same value; this means that for the first axis, all experimental conditions have the same weight, in other words, the first axis gives us the total expression of each gene, just like with a spreadsheet. This observation is generally true (see Stoyanova et al., 2004).

In each of the other columns (axes), the experimental conditions can be grouped together according to the sign of their coefficient (positive or negative). For some axes, this coincides with a separation of the two states of a factor. In our case, axis two separates well the two protocols (1 μg RNA: all values are negative and 10 μg RNA: all values are positive); axis three separates the day (A and B), axis five the sulphur source (met and mtr) and axis seven the two spots (a and b). Other axes, on the other hand correspond to combinations of the experimental conditions, whose interpretation is not evident: axis four is an example. It singles out the ribosomal proteins; a biologically speaking coherent result, which is waiting for an interpretation. This is something frequently found when analyzing microarray data.

The *eigenvector matrix* deserves a little more attention: the values it contains can be looked at from a different point of view. If we take up our example, each line represents an experimental condition and the values in the 16 columns for a given line give us the position of that particular experimental condition in the 16-dimensional space. (To be precise, each value has to be multiplied with the root of the variance of

Table 1 (*Continued*)

The mixing matrix at the top was calculated by the spreadsheet, the mixing matrix (or *eigenvector matrix*) at the bottom by PCA. The arrows indicate the columns which separate well the effects of the same factors. The matrices allow us to change from the old to the new reference frame: they give us for each of the new axes (the columns) the coefiicient with which we have to multiply each gene's value in a given experimental condition (the lines) in order to obtain the new coordinates. The first line in the *eigenvector matrix* contains the *eigenvalue* for each axis (in %), providing an indication of the cloud's dispersion along that axis. Note that for the first axis all the sixteen coefficients have basically the same value; this means that for the first axis, all experimental conditions have the same weight, in other words, the first axis gives us the total expression of each gene, which is generally true (see Stoyanova et al., 2004). An example for the calculation of the new coordinates with the *eigenvector matrix*: in the original (or "old") reference frame, the gene *galK* has the coordinates (5.431; 5.432; 5.092; 5.068; 4.893; 4.744; 3.763; 3.661; 5.333; 5.265; 5.329; 5.249; 4.607; 4.444; 3.806; 3.737). To obtain *galK*'s coordinate on the new axis 1. the calculations are as follows: $(5.431 \times 0.250) + (5.432 \times 0.249) + \cdots + (3.737 \times 0.251) = 19.0$. The other coordinates are obtained accordingly.

that axis, in order to obtain the coordinate.) This means that instead of looking at the *eigenvector matrix*, we can look at the different projections of the experimental conditions in order to figure out which axes separate well the different states of our factors. Once we have established which planes deserve being examined in details, we come back to the projections of the cloud on these planes and pinpoint those genes, which are far away from the main body of the cloud. Fig. 5 shows the cloud projection on the plane formed by axis one versus axis five.

Note: the normalization of the data is an integral part of PCA.

To resume, with PCA the experimental conditions are not given the same weight (contrary to a spreadsheet) and the criterion chosen to look at the cloud is to maximize the variances along the axes.

3.3.4. ICA

"ICA tries to find a linear representation of non-Gaussian data so that the components (or factors, or sources) are statistically independent, or as independent as possible" (Hyvärinen and Oja, 2000).

This search for statistical independence is generally very difficult and therefore an approximation is made: one looks for the directions that maximize the criterion of non-Gaussian distribution. As "non-Gaussian" is a "non-property", numerous possibilities exist for defining such a distribution. One criterion that seems to work quite well is to look for distributions with a positive kurtosis (distributions with "heavy tails"). ICA can be seen as a close relative of PCA. Whilst PCA looks at which directions maximize the variance, ICA approaches the question of finding genes with an "atypical behaviour" more directly, by defining "atypical" as "following a non-Gaussian distribution". The new reference frame will maximize the criterion of "non-Gaussianity". With this criterion, one increases the weight of points that had only small deviations from the main body of the cloud and thus allows them to be detected as potentially interesting.

A latent difficulty with ICA is that there is no analytical solution (contrary to PCA): we look for the numerical solutions. There is the danger that the algorithm finds a direction with a solution, but that this direction is not the best solution

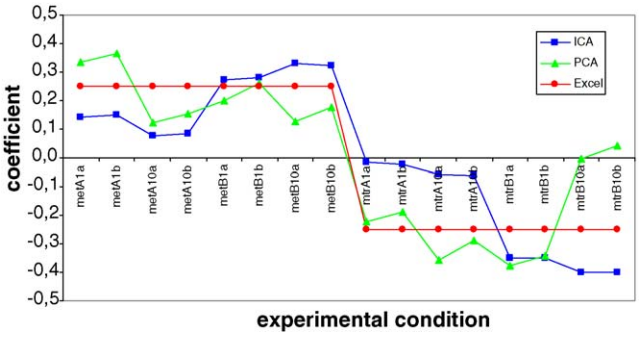


Fig. 7. The "weight" attributed to each experimental condition by a spreadsheet (MS-Excel), PCA and ICA. The figure shows that a spreadsheet (MS-Excel) attributes the same "weight", or importance, to each experimental condition, whilst PCA and ICA do not.

in absolute terms: the algorithm gets stuck with a local maximum (Chiappetta et al., 2004). Launching ICA a large number of times, typically 100, circumvents this problem and only those directions or solutions that have been most frequently found are kept. As with PCA we have a mixing matrix that allows us to change from the old reference frame to the new one. Again, the different experimental conditions do not have the same weight; the weight attributed, though, varies slightly from PCA. Once we have determined the axes, the procedure is the same as with PCA. Fig. 6 shows the cloud projection on the plane that separates well the sulphur sources.

The applications of ICA in microarray analysis include the identification of groups of genes implicated in cancer, the study of the cell cycle (Liebermeister, 2002; Martoglio et al., 2002) and the identification of genes that are potentially co-regulated (Chiappetta et al., 2004). Chiappetta et al. (2004) and Carpentier et al. (2004) have applied both PCA and ICA to the sulphur metabolism data and shown that the two methods perform similarly well, with ICA slightly outperforming PCA.

3.3.5. A brief remark

We have said that whilst a spreadsheet attributes to each experimental condition the same weight, PCA and ICA do not (Fig. 7 shows a comparison between the three methods from this point of view).

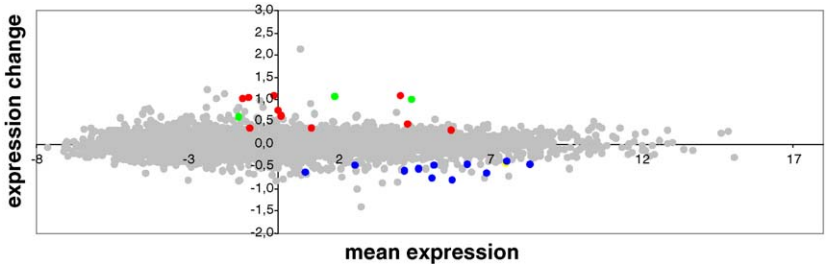


Fig. 6. The data cloud projected on the plane that separates well the sulphur source (ICA). The figure shows the genes' expression change in function of the sulphur source, as determined by ICA. The potentially interesting genes are those away from the main body of the cloud. The highlighted genes are the ones which proved to be of particular interest for the problem investigated by Sekowska et al. (2001); the reader may refer to their work for a detailed discussion. Note that not all of these genes would have been detected using ICA.

The spreadsheet simply calculates the mean expression. This choice is not optimal when certain experimental conditions contain more information than others. Ideally, the weight attributed should be proportional to the information contained. PCA is a good choice when the signal follows a Gaussian distribution, whilst ICA imposes itself when the distribution is non-Gaussian.

You might wonder what happens if you use, say, PCA on data that follow a non-Gaussian distribution. The answer is that you are likely to miss out on potentially interesting genes; you do not, however, risk finding "wrong" genes. Using more than one tool amounts to examining the cloud from different angles; the results obtained with the different tools are complementary.

3.4. Statistical tests

3.4.1. Preliminary considerations

Our experience shows that some confusion reigns regarding the statistical tools in general and their application to microarrays in particular. Hence this rather long introductory section.

When approaching microarray data from a statistical point of view, people seem to worry a lot about the fact that the data are "relative" and whether they should or not take ratios.

Microarrays give us "relative data": the interesting information regarding a gene is "relative" as one compares the expression of a gene under condition A with that of the same gene under condition B. Microarray technology is quite recent; however, dealing with relative data is not and taking the ratio results in a reduction and a falsification of the information offered (Kerr and Churchill, 2001). It is Fisher who first tackled and solved the problem at the very beginning of the 20th century, resulting in ANOVA. For a more detailed discussion of this issue, the reader is referred to the work of Kerr and Churchill (2001). At about the same time, Gosset ("Student") came up with the *t*-test as a solution to the problem.

Statistics help us to answer the question whether the expression differences observed are real. The answer is given indirectly, as the statistical tools give us the probability of having a false positive. A false positive is a gene whose expression difference surpasses by chance a threshold value, which has been fixed in advance. "By chance" means that if the experiment were repeated, you would not find again such a large expression change.

The statistical analysis is used to evaluate the probable percentage of false positives beyond a given threshold value: 40 genes will surpass by chance the threshold value of 1% if the experiment was carried out on 4000 genes.

The estimation of the number of false positives is only the first step. Beyond the threshold value we not only find false positives, but also genes whose expression change is "real" (we would find it again if the experiment were repeated). The key information is the proportion of false positives on the total, because it measures the risk of being on the wrong track when deciding to work on one of the genes from this group (Benjamini and Hochberg, 1995). One generally chooses the threshold in order to have less than 5% of false positives in the group. Take for example an experiment carried out on 4000 genes with 80 lying beyond the threshold of 0.1%. As there are on average four false positives beyond the 0.1% threshold (4000×0.001), the percentage of false positives is 4/80, or 5% of the selected genes.

The literature sometimes refers to the Bonferroni correction. This correction is not pertinent for the analysis of microarray data, as it is too restrictive.

The numerical criterion used in the statistical tests is always the ratio between the deviations observed for the factor of interest (the signal) and the deviations due to all the causes one chooses to ignore (the noise). The statistical tests differ from each other in the way they define the noise and the probability function they use to estimate the probability of false positives. In the past, the function used was the Gaussian. Nowadays one tends to employ the probability function, estimated on the data using permutations (see Tusher et al., 2001).

3.4.2. ANOVA

ANOVA is a tool that allows us to analyze simultaneously the effect of more than one factor on a variable, in our case the genes' expression levels. The method is based on the calculation of the sum of squares, degrees of freedom, mean square (short for mean square deviation from the mean) and *F*-statistics[1] (see Zar (1998) for details). As we use ANOVA in a somewhat reductive manner, the reader may refer to the work of Zar (1998) for a full appreciation and pedagogic explanation of the possibilities offered.

Various quantities are used simultaneously in order to decide whether the expression of a gene varies significantly for the factor of interest.

1. V1, the variance for the total of the observations made on the gene;
2. V2, the variance for the observations made for the factor of interest;
3. V3, the variance for the observations made for those factors whose influence one wishes to subtract.

The signal is equal to V2, the noise to $V1 - (V2 + V3)$. The possibility to calculate the term V3 is a particularity of ANOVA and it allows a finer control of the noise's composition. In our example, V3 corresponds to the expression change caused by the day, the duplicate and the RNA concentration. The noise encompasses all which causes the difference between the actual expression level and the sum of the expression levels of the four factors.

In the case of the sulphur metabolism data, the equation used for each gene is the following:

$$Y_{ijkl} = \mu + S_i + J_j + C_k + D_l + \varepsilon_{ijkl}$$

[1] Sometimes referred to as *F*-test.

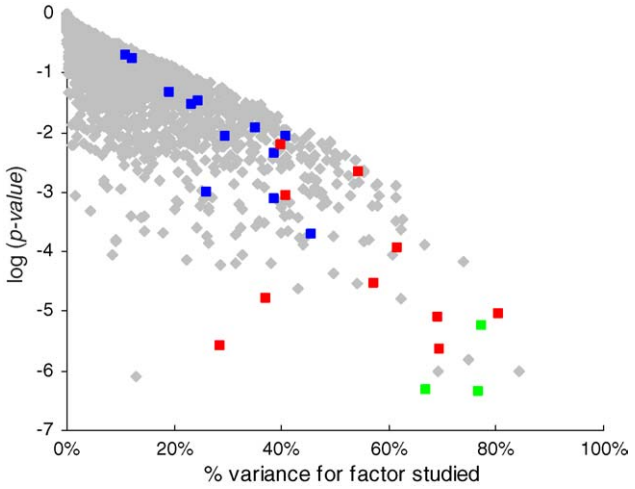


Fig. 8. The graphical representation of the results obtained with ANOVA for the factor sulphur. The potentially interesting genes are those with a small *p*-value and a large variance, genes which are therefore in the bottom right part of the image. They are away from the main body of the cloud. The highlighted genes are the ones which proved to be of particular interest for the problem investigated by Sekowska et al. (2001); the reader may refer to their work for a detailed discussion. As can be seen, not all genes of interest would have been identified by the sole use of ANOVA.

where Y_{ijkl} is the total expression level measured; μ the mean of the expression levels measured for the gene; S_i, J_j, C_k and D_l, respectively, the effects of sulphur source i, experiment day j, RNA concentration k and duplicate l on the expression level; ε_{ijkl} is the residual error.

Note that the residual error ε_{ijkl} encompasses all interactions: between two factors (6), between three factors (4) and between four factors (1). The interactions are grouped together under "error" for the following reason: it is information with which we cannot work, unless we have a very precise idea of the nature of the interaction (linear, sinusoidal or other).

The *F*-test is calculated in the following manner: *F* = "mean square of the sulphur source"/"mean square of the residual error". We are interested in genes that posses a high *F*-value (*p*-value) for the factor sulphur source. The calculations are done for all genes and the results can be represented in a graphical form. The variance of the factor of interest is given on the *x*-axis, the *p*-value on the *y*-axis. The *p*-value is used to calculate how many false positives will lie below a chosen threshold value (see Fig. 8).

Note that we are not interested, whether the expression levels of the thus identified genes also vary in function of the other factors. One does not preclude the other and has no impact on our analysis.

ANOVA has some advantages when the experimental factors are not binary; in that case, it basically becomes the only tool which is easy to use.

3.4.3. Paired t-test

We have said that ANOVA quantifies the contribution given by each factor to the total expression of a gene, permitting us to isolate the contribution of our factor of interest. The paired *t*-test also allows this, but the approach is different, and we can only use it for binary factors. The paired *t*-test eliminates the influence of all the factors we are not interested in by calculating the difference between pairs of values. The members of each pair differ from each other only with respect to the factor of interest (state 1 versus state 2), all other experimental conditions being equal.

For example, we calculate the difference between the value obtained on met with the value obtained on mtr, both obtained on day A, with 1 μg mRNA and spot a. Then we calculate the difference of met versus mtr on day B, with 1 μg mRNA and spot a and so forth. This is done for each gene and we thus obtain eight comparisons, or differences per gene. V1 is calculated on these eight comparisons, the term V3 has disappeared.

However, as the paired *t*-test takes pairs of "similar conditions", systemic biases due to, e.g. "day" or "duplicate", are eliminated, therefore still allowing for a reasonable estimation of the error.

3.4.4. t-Test

The *t*-test corresponds to an ANOVA with one factor and is the least favourable option. The *t*-test only considers the expression difference due to one factor, ignoring that there are pairs of measurements which have more or less in common (like the day, protocol and spot), unlike ANOVA and the paired *t*-test. Thus, we cannot separate the contribution made by our factor of interest from the contribution made by the other factors and the interaction between them; the expression difference due to our factor risks being drowned by the rest.

In terms of V1, V2 and V3: V1 is calculated on the total of the 16 observations made (as with ANOVA), but as the term V3 has disappeared, the noise risks being much larger.

3.4.5. In conclusion

The biggest difficulty is to estimate the noise with accuracy. The best solution is to repeat the experiment a large number of times. As this is not always possible, statisticians try to improve the estimation of the noise by working on groups of genes having more or less the same level of noise. A considerable amount of literature is dedicated to this effort. Numerous are the solutions proposed, none is perfect. Generally, the grouping is done a posteriori, after a first estimation of the noise for all the genes separately. One speaks in this case of a Bayesian approach. The reader is referred to the work of Neuhäuser and Senske (2004) for an introduction into the subject and to the work of Kutalik et al. (2004) for the comparison of some methods proposed.

Regarding the three approaches discussed above: Table 2 shows the measurements obtained for *ytmJ* and the results obtained from ANOVA, the paired *t*-test and the *t*-test. It shows that though ANOVA and the paired *t*-test both identify the gene as interesting, the *t*-test results inconclusive. The observation made on this particular example can be

Table 2
Comparison between ANOVA, the paired *t*-test and the *t*-test, an example

	met	mtr	met–mtr
(a) Measurements obtained for *ytmJ*			
A1a	1.170	1.520	−0.3494
A1b	1.176	1.580	−0.4048
B1a	0.950	1.566	−0.6158
B1b	0.891	1.541	−0.6496
A10a	1.939	2.049	−0.1096
A10b	1.565	2.048	−0.4827
B10a	0.893	1.523	−0.6296
B10b	1.007	1.485	−0.4772
(b) *t*-Test			
Numerator	−0.465		
Denominator	0.313		
d.f.	14		
t-Test	−2.971		
−log(*p*-value)	1.99		
(c) Paired *t*-test			
Numerator	−0.465		
Denominator	0.064		
d.f.	7		
Paired *t*-test	−7.290		
−log(*p*-value)	3.78		

	State 1	State 2	SS
(d) ANOVA			
Sulphur	9.592	13.311	0.864
Day	13.048	9.855	0.637
RNA	10.394	12.508	0.279
Spot	11.610	11.293	0.006
Residual			0.448
Total			2.235

Factor	SS	d.f.	Variance	*F*	−log(*p*-value)
Sulphur	0.864	1	0.864	21.21	3.12
Day	0.637	1	0.637	15.64	2.65
RNA	0.279	1	0.279	6.86	1.62
Spot	0.006	1	0.006	0.15	0.00
Residual			0.041		
Total			0.149		

In (a) the measurements obtained for *ytmJ* are shown. (b–d) The calculations and results obtained with the *t*-test, the paired *t*-test and ANOVA, respectively. ANOVA and the paired *t*-test both identify the gene as potentially interesting, whilst the *t*-test results inconclusive (see the relative −log(*p*-value)). d.f. = degrees of freedom; SS = sum of squares. See Section 3.4 for details.

generalized. Table 3 shows a comparison of the number of genes detected by the three methods. Although ANOVA detects the highest number of genes, the paired *t*-test performs comparably well, whilst the *t*-test lags far behind.

3.5. Graphic exploration and statistical tests in comparison

We have chosen to talk about the typical representatives of the two approaches. They are not the only ones proposed in the literature: the number of tools is continuously increasing and no one, definitive method has so far emerged, as is exemplified by the web-site maintained by Li, which has a steadily growing collection of articles on microarray data analysis (http://www.nslij-genetics.org/microarray/). Conceptually, all these tools are based on one of the methods described above or they fall into the category “cluster analysis”, described below.

Some methods will use the term “distance”, whilst others may talk about “correlation”. In mathematical terms, it boils down to the same thing: second order statistics, yielding the same type of information. As the methods all differ more or less from each other, it is normal that they do not come up with exactly the same results.

Which method is the best? Carpentier et al. (2004) have examined this issue and developed a protocol that allows the comparison of the different methods, in terms of their reliability. They conclude that each of the methods analyzed gave

Table 3
Overall comparison between ANOVA, the paired *t*-test and the *t*-test

(a) Number of genes detected with a threshold of $-\log(p\text{-value}) = 3$	
ANOVA only	62
t-Test only	5
Both	24
ANOVA only	35
Paired *t*-test only	30
Both	51
t-Test only	12
Paired *t*-test only	64
Both	17
Total ANOVA	86
Total paired *t*-test	81
Total *t*-test	29
(b) Number of genes detected with a threshold of $-\log(p\text{-value}) = 4$	
ANOVA only	16
t-Test only	0
Both	9
ANOVA only	17
Paired *t*-test only	12
Both	8
t-Test only	14
Paired *t*-test only	15
Both	5
Total ANOVA	25
Total paired *t*-test	20
Total *t*-test	9

The table shows a comparison of the number of genes detected by the three methods. In (a) the threshold for detection was $-\log(p\text{-value}) = 3$, in (b) it was equal to 4. Although ANOVA detects in both cases the highest number of genes, the paired *t*-test performs comparably well, whilst the *t*-test lags far behind.

some information not provided by the others, suggesting once more the advantage of analyzing one's data with more than one statistical tool.

ANOVA, one of the methods tested, did not excel on the sulphur metabolism data. However, all factors were binary and ANOVA has the great advantage of being easily applicable in cases where the factors are non-binary. It also has another important property: ANOVA is the only method that forces the experimenter from the beginning to give the experimental setup some thought, to plan it carefully. It is therefore a good practice to think of an experimental setup in terms of ANOVA, even if the data are then exploited by another method (see Section 4.2).

3.6. And the clustering approach?

The principle is to group and/or to classify the genes in function of the expression profile obtained under the various experimental conditions.

The cloud is thus divided into a number of clusters, the idea being that a cluster corresponds to a functional class. Choosing a gene of unknown function, one can look to which cluster it belongs and thus draw conclusions about its possible role.

This approach poses problems from two points of views: a biological and a technical one.

From a biological point of view: we have to define what a functional class is and how many there are. These are not banal questions, as exemplified by the fact that even for such a well-studied organism like *E. coli* numerous classifications are proposed (for example SwissProt, EcoCyc, Kegg). Secondly, the functional classes found in the literature tend to be rather large, containing dozens or hundreds of genes, making them too large to permit their exploitation in the wet lab. Thirdly, the clustering methods normally do not allow a gene to be part of more than one cluster, which goes against biological intuition and experience.

From a technical point of view: we have to choose amongst a myriad of (family of) clustering techniques. As the biological question is not clearly defined, we do not have a criterion to select the pertinent and coherent method for our needs.[2] At this point one has to make do with a data-driven attitude. This necessitates a thorough knowledge of the different families of clustering techniques in order to make the best choice in function of the data set to be analyzed (Somorjai et al., 2003), as all the clustering techniques require many prior decisions (Chiappetta et al., 2004). In addition, as Somorjai et al. (2003) point out: "the maxim 'simpler is better' has mostly been ignored".

As clustering methods are well-liked tools (see for example the popular software proposed by Eisen et al. (1998)), various attempts have been made to circumvent the various technical problems. The reader, who would like to have a critical introduction to different families of clustering techniques, may refer to the works of Datta and Datta (2003), De Smet et al. (2002) and Somorjai et al. (2003).

3.7. And SOM?

Generally speaking, only the outskirts of the cloud are visually exploitable. The internal organization is hidden by the superposition of thousands of genes on the same image. The analysis would be easier if it were possible to give a faithful representation of the genes' density in each region of the cloud, with only *k* points. A rather naïve solution consists in choosing these *k* genes at random. This is unlikely to give satisfactory results, though. Calculating the optimal position of the *k* points is a difficult problem. A number of programmes exists proposing approximate solutions. An example is self organizing maps (SOM), which chooses the *k* genes and provides a list of the genes close to the *k* genes. The interested reader may refer to the work by Kaski et al. (2003) for an

[2] An example is the definition of the distance between clusters. This is not a banal problem. Take for example the problem of having to define the distance between two countries: do you take the two capitals? The two biggest cities? The shortest distance (0 if the countries are adjoining)?

introduction to SOM as well as a comparison of its merits compared to some classic classification methods.

Note that all the programmes proposed necessitate the adjustment of numerous parameters for which you do not necessarily have a rational basis to make your choice. This carries the risk that you only believe those results which tell you something you already know: not the best way to discover new things.

4. Intricacies of microarray-based methods

4.1. FAQ

Over the years of teaching the course on the analysis of microarray data, we have noticed that certain questions, more or less closely related to the subject, turn up on a regular basis. Here are some of them, with the answers.

4.1.1. Missing values

They have generally two possible origins: (i) the microarray contains a defect resulting in the impossibility of taking a reading or (ii) the machine eliminates the measurement as the value is very close to the noise level (in this case it would be advisable to change the setup of the machine). This poses a problem, as many data analysis methods require full sets of data. The most radical solution is to eliminate the genes with missing entries, which is obviously far from ideal. A more moderate solution is to fill in the gaps with estimate values. The easiest is to use the row average; the two most common methods, however, are:

(a) looking with whom the "missing gene" associates with in the other experimental conditions, i.e. determining that gene's "neighbours", then presuming that in the missing experiment this gene still associates with them and filling the gap with the median value (a method known under "K-nearest neighbours");
(b) variations around the PCA (examples here are the singular value decomposition and Bayesian principal component analysis (Oba et al., 2003)).

The interested reader is referred to the works of Ouyang et al. (2004), Kim et al. (2004) and Zhou et al. (2003) for the comparison of some currently used estimation methods.

4.1.2. The correction of the background noise on the membranes, glass plates or silicon chips

This problem tends to be given too much importance. ANOVA allows us to easily quantify the inter-array variation and the result is that this variation is small compared to other sources (for example "day" in the case of the sulphur metabolism data, see Chen et al. (2004) for a detailed discussion on this subject), strongly suggesting that the effort spent correcting background noise is not justified. There is also a second aspect to be considered, namely that finding a reliable method to correcting background noise is not easy. As Lawrence et al. (2004) point out, the human component plays an important role. The assumption that the background level is consistent between the DNA spot and the surrounding space, frequently used for background quantification, is not correct (Konishi, 2004). Using "designated" household genes for the background determination is in itself a good idea, but finding out who the household genes are, is posing problems (Stoyanova et al., 2004).

Regarding Affymetrix's GeneChips, the common practice of subtracting the mismatch (MM) probe intensities from the perfect match (PM) ones is "unjustifiable", according to Sasik et al. (2002), as the target sequence hybridizes not only with the PM but also with the MM probe.

We remind the reader at this point that a useful way to assess the utility of an anti-noise measure taken is to check on the change of the *eigenvalue* of the first axis in a PCA (should increase) or the *F*-value in an ANOVA (should increase).

4.1.3. Dealing with data containing a large number of very small or zero values

You may find yourself in the situation of not having access to the "real data": you are given a set of data, where all values below a certain threshold were replaced with one or very few arbitrary values. This means that the distribution is far from being Gaussian (or just unimodal), a fundamental prerequisite for the analysis of the microarray data.

The only solution to this dilemma is to try to "restore" the Gaussian distribution by replacing the smallest values with random values (see Chiappetta et al., 2004).

Having a large number of very small or zero values may simply be the result of a translation, usually the effect of having subtracted the background noise. In this case, it is sufficient to add to all values a constant (for example the weakest signal measured in the experiment) before taking the log of the data.

4.1.4. Taking the ratio or not?

Microarrays give us "relative data": the interesting information regarding a gene is "relative" as one compares the expression of a gene under condition A with that of the same gene under condition B. Microarray technology is quite recent; however, dealing with relative data is not. The following text is taken from Kerr and Churchill (2001) who discuss the issue in a very clear and lucid manner: ". . . relative data is about as old as statistics itself. The "grandfather" of statistics, R.A. Fisher, worked with agricultural field trials. In controlled experiments with clear objectives, scientists sought to determine the productivity of different varieties of a crop, for example different strains. They recognized that there is no such thing in absolute terms as the yield of a variety because productivity depends on soil fertility, sunlight, rainfall, and myriad other factors. They understood that the only meaningful direct comparisons are for strains grown on the same block of land. Consider a hypothetical experiment to study three varieties. Suppose there are three blocks of land available, but each block only has room for two varieties. . . ..

It is easily accepted that the yield data contain information about the varieties grown in the same block. However, there is a corresponding fact relying on the same logic that can be overlooked. Namely, there is also information about the blocks of land because they have varieties in common. Fisher recognized this duality and realized one could simultaneously estimate the relative yield of varieties and the relative effects of the blocks of land. The quantitative tool for doing this is a simple linear model:

$$y_{ij} = \mu + B_i + V_j + \varepsilon_j$$

where y_{ij} is the measured yield for variety j grown on block i; μ the overall mean; the block effect B_i is the effect of block i; and V_j is the effect of variety j. The term ε_{ij} represents random error. In a large experiment with many varieties and blocks, unbiased yield comparisons can be made, even for varieties not grown on the same block of land. Returning to microarrays, consider the spots for a particular gene on different arrays (or reproduced within arrays). The spots vary in size, shape, and concentration, analogous to the variation in fertility of blocks of land. Using the same principles as in the agricultural experiment, we can simultaneously measure the relative transcription level of the corresponding gene and the "fertility" of the spots. However, this is only possible if we use all the information in the data and do not reduce to ratios."

This should answer the question adequately.

4.1.5. The problem posed by the two fluorescent dyes used with glass plates

When working with glass plates, one is given the choice between two different dyes to be used for the incorporation in order to obtain the hybridization mixture. In other words, one is given the choice between two experimental protocols. In the case of the sulphur metabolism experiment the choice was between using 1 or 10 μg mRNA. When ordering the genes as a function of the average intensity of the signal, Sekowska et al. (2001) observed that the order is highly sensitive to the protocol; on the other hand, if the same protocol is used, the results are highly repeatable (see Tables 5 and 6 in Sekowska et al., 2001). This observation can be generalized.

Does this have an impact for the analysis? The factor protocol is an important source of variability (Chen et al., 2004; Sekowska et al., 2001), but as there is no interaction between the other factors, the impact on the analysis is minimal with all the techniques described here (PCA, ICA, ANOVA and paired *t*-test). Having two protocols, doubles the number of measurements without being instructive on the biological problem studied. If one has the means (economic or other) to increase the number of measurements, it is perhaps advisable to introduce a biological repetition or a new biological factor (Chen et al., 2004).

4.1.6. How many genes should I put on my microarray?

The answer is simple: as many as possible. The reason is the following: a number of analysis methods normalize the data, which is only justified if three conditions are fulfilled (see "The spread sheet and some preliminary considerations and manipulations" above): firstly, more than 90% of the genes do not change expression in function of the different experimental conditions, secondly, the number of genes analyzed is large and thirdly the overall intensity change of up- and down-regulated genes is similar (see Stoyanova et al., 2004).

4.1.7. What can I do if my signal is outside the linear range (of my machine)?

This results in a "cigar" which is twisted and bent. The first and obvious recommendation is to make sure that at the moment of taking the readings, the scanning settings are correct, which they are often not (Stoyanova et al., 2004). The second is to check that one is not just working at one extreme of the linear range; if that is the case, a change of concentration in the hybridization solution is a good option. If the entire linear range is taken up, two solutions can be proposed: using two different voltage settings for the photomultiplier or using different exposure times, when working with radioactively labelled samples. Algorithms for subsequently combining the different readings are readily available (see for example Querec et al., 2004; Lyng et al., 2004). The article by Lyng et al. (2004) shows the relationship between the type of incorrect setting and the resulting cloud shape.

Numerous authors propose "remedies" if the above suggestions prove impossible to follow, but none will give you the "perfect" data back you would have had if the experiments had been executed correctly.

4.1.8. How does one tackle a temporal series?

In time series expression experiments a number of samples is taken over a period of time. Biological and computational problems specific to this type of experiment have to be faced from the experimental setup to the data analysis and to the interpretation of the data. The reader is referred to the work by Bar-Joseph (2004) who reviews these problems and the solutions offered.

4.1.9. How do we find genes for an accurate diagnosis of a disease?

Typically, the data will come from one hospital and from a relatively small number of patients. These patients represent the learning set and the analysis of the data will always come up with some candidate genes. To validate the results, however, we need a validation set. It is wise to have five to six times more patients in this set than candidate genes. To avoid finding genes that are only specific to a particular socio-cultural-genetic background, the patients should be chosen from more than one hospital and more than one country (see also Section 4.2.1).

From a theoretical point of view, the use of microarrays for the diagnosis of a diseases poses two fundamental problems, the first one being Bellman's "curse of dimensionality" (too many features or dimensions, e.g. thousands of genes), the

second one being the "curse of dataset sparsity" (too few samples) (Somorjai et al., 2003); this means that we end up analyzing a space with a great number of dimensions which is nearly empty: whatever method is applied to the analysis of the data, the result is unlikely to be statistically sound, the biological interpretation risks being inconclusive.

Somorjai et al. (2003) discuss this problem in detail. Hwang et al. (2002) propose a power analysis method in order to determine the minimum sample size for the – statistically reliable – discrimination of distinct disease states.

4.1.10. How do we determine the relative importance of a factor?

By using an ANOVA, as it explicitly estimates the magnitude of the sources of variation and therefore gives us the relative importance of each factor (see also Chen et al., 2004).

4.1.11. What does the p-value tell me? What about false positives?

The *p*-value gives us the probability of finding by chance a deviation from the mean equal to or larger than the one we observe.

For example, if we decide to work on all those genes with a *p*-value smaller or equal to one per mille (0.1%) and we examine 4000 genes, we expect on average 4 genes to fulfil this criterion by chance, without reflecting a biological reality. These four genes fall into the category of false positives. With a *p*-value equal to 5%, we would expect 200 genes to fulfil the criterion by chance.

This means that the *p*-value helps us to judge and quantify the risk of looking at or working with a false positive, nothing more and nothing less.

4.1.12. Will I not miss out on a few genes?

Through the years of teaching, we have noticed that this is apparently a worry common to a lot of people working on microarrays.

Even provided that the experiments were planned and executed in a diligent manner, the problem will generally not be that of having too little genes changing expression, but too many. An example is given again by the sulphur metabolism experiment where less than a dozen genes were of actual interest, truly involved in the phenomenon studied, but a lot more changed expression. The reason for this "surplus" of genes is that there will always be secondary effects. Biological processes seldom come in a straight line; more often they resemble an intricate net (think of the cell's metabolism), which means it is near impossible to isolate a phenomenon completely (see Sontag et al., 2004). To further narrow down the list of candidate genes, one will have to use any available (biological) knowledge from other sources.

4.2. How to plan one's experiment?

It is quite usual to find that a rather large number of genes, typically around 10%, change expression considerably between two experimental conditions. This number is too large to be directly exploitable and we will have to extract a short and pertinent list of genes to work with. This task is greatly facilitated by an adequate and well thought-through experimental setup.

4.2.1. The type of factors

An experiment is made up of three types of factors, each providing specific information.

The first factor corresponds to the phenomenon studied. The study concern two or more states (two culture conditions, for example, or a certain number of samples taken during a time course experiment). The aim is to narrow down to a maximum the target genes, in other words to have only few genes who change expression considerably between the different experimental states. For this, the experimental states should be as close as possible, for example:

(a) In the case of the sulphur metabolism experiments, the two sulphur sources were metabolically speaking closely related.
(b) When trying to isolate genes typical of a certain cancer, one should study different subtypes, all closely related to the one of interest.

If this maxim is not observed, too many genes will change expression considerably and the identification of target genes will become near impossible.

The second type of factor serves to verify whether the observations made hold true if the biological parameters are changed. Do we find the same candidate genes if we work with a different bacterial strain? Or patients from a different hospital? Or if the experiment is carried out on a different day? Note that even repeating the experiment on a different date introduces a biological variability, as the experimental conditions will never be exactly the same (see Sekowska et al., 2001). This verification is extremely important as the most interesting genes are those which come up whatever the biological parameters. They are most likely the genes at the heart of the phenomenon studied, as their behaviour is not bound to a particular context (genetic, socio-cultural or physiological). The reader is referred to the works of Turk et al. (2004) and Whitehead and Crawford (2005) who discuss this issue.

The third type of factor is a technical one: the type of protocol used to label the cDNA, having two spots for each gene on the array or using the dye swap. This type of factor increases the workload without adding any biologically pertinent information. The experimental protocols have become highly reproducible and it is advisable to stick to just one protocol (with its systemic biases) and increase the number of states of the two other factors.

4.2.2. The ideal situation: a fully crossed factorial design

The best experimental setup is to follow a fully crossed experimental design (exemplified by ANOVA) as it

(a) allows a good exploitation of the information given;
(b) allows a precise estimation of the error variance (see Fisher (1951) for the original discussion or Mather (1943), Zar (1998) and Kerr et al. (2000) for a more user-friendly approach).

Setting up a fully crossed factorial design means that each level (state) of one factor is found in combination with each level of the other factors, as shown in Fig. 1. Note that carrying out twice the experiment on strain 1 on day A and twice the experiment on strain 2 on day B would not be adequate as it would be impossible to separate the effect of the day from the effect of the strain.

4.2.3. The reality

A fully crossed factorial design may not be possible. This is typically the case in clinical studies, as they strongly depend on the hospitals' random recruitment of patients. These represent a learning set. To confirm the results a validation set would be needed, and to avoid finding genes specific to a particular socio-cultural-genetic background only, the recruitment should be made at more than one hospital and more than one country.

This may prove to be unfeasible if not impossible. A different option is to give up on the fully crossed factorial design altogether and take a completely different approach: one can exploit all the experimental data available in the literature (freely available on the web) by pooling them together. This is not as bizarre an idea as it may seem; the aim can be to increase the number of patients (Jiang et al., 2004) or to get information about the co-expression and co-regulation of genes (Lee et al., 2004; Yeung et al., 2004). Especially for the two latter issues, this is the only approach: as a very large amount of data is needed, which a single lab could not possibly come up with.

Various authors propose statistical models to help extracting the maximum information from these pooled data (see for example Shen et al. (2004) and Statnikov et al. (2005)). Note that when working with pooled date, their analysis will have to be carried out with methods which do not need the definition of the factors a priori, like PCA or ICA.

4.2.4. The combination of factors

If we want to obtain useful information from our microarray experiment, we are forced to formulate precise questions. This means that we cannot combine two factors in one question, as this is equivalent to measuring the interaction between the factors, which is not separable from the error (unless we know in detail the relationship between the two factors, for example linear or sinusoidal).

This is one more reason to follow a fully crossed factorial experimental setup, as exemplified by ANOVA. It forces us to spell out in detail what we want to measure and what will be part of the error or interaction component. It has the great advantage of permitting the identification of the sources of variability and their magnitude; this allows making the improvements to the setup at the right sources, which will generally be a modification of the experimental protocol and an increase in the number of biological replications (Chen et al., 2004).

5. The answer to all our questions?

Microarrays are sometimes seen as the miracle tool, which will give all the answers to all the questions. Paying considerable attention to the experimental setup is a necessary condition, but not a sufficient one.

The preliminary phase should already take into consideration the different analysis options available to the experimenter by pulling in statisticians. This should be an exchange, not a handing over the job to a statistician, as the biological question has to stay at the front. As Vingron (2001) points out in his editorial, bioinformaticians should "go back to school and learn more statistics. Not so much with the goal of mastering all of statistics but with the goal of sufficiently educating ourselves on order to pull in statisticians."

A careful analysis of the data should follow (we suggest using more than one method, as they tend to give complementary information).

The complex nature of biological phenomena means that it is near impossible to isolate the candidate genes only through a microarray experiment (Curtis and Brand, 2004; Somorjai et al., 2003; Sontag et al., 2004), meaning that the list of candidate genes obtained will have to be further worked on. This may be done by further theoretical work (integration of all available biological knowledge) or additional experiments in the wet-lab.

Sometimes the a priori biological knowledge about the phenomenon of interest may be very limited. In this case, it can happen that despite careful planning and execution, the genes identified as interesting are not actually the cause of the phenomenon. This was for example the case in the genome-wide analysis undertaken by Oshima et al. (2002). In these cases a new experimental setup may be solution. However, the conclusion may also be that a transcriptome analysis is not the adequate tool for the study of the phenomenon (Riva et al., 2004).

It is therefore important to realize that a microarray analysis will generally not be THE answer to all your questions. It is a complement to other approaches.

6. Software and data used

The sulphur metabolism data from Sekowska et al. (2001) are freely accessible at http://195.221.65.10:1234/~carpenti/.

PCA and ANOVA were performed using GeneANOVA, freely available on request for non-commercial use. Please contact Gilles Didier at didier@iml.univ-mrs.fr.

ICA was adapted to gene expression analysis by Bruno Torrésani, Pierre Chiappetta and Marie-Christine Roubaud (see http://www.cmi.univ-mrs.fr/~torresan/publi.html).

Acknowledgments

The authors wish to thank J.-L. Risler for critical reading of the manuscript. This work was supported in part by the French Ministry of the Economy, Finance and Industry (contract ASG no. 01 4 90 6093).

References

Bar-Joseph, Z., 2004. Analyzing time series gene expression data. Bioinformatics 20, 2493–2503.

Barrett, J.C., Kawasaki, E.S., 2003. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discov. Today 8, 134–141.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc., Ser. B 57, 289–300.

Bussemaker, H.J., Li, H., Siggia, E.D., 2000. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc. Natl. Acad. Sci. U.S.A. 97, 10096–10100.

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., Rubin, E.M., 2000. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res. 10, 2022–2029.

Carpentier, A.-S., Riva, A., Tisseur, P., Didier, G., Hénaut, A., 2004. The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. Comput. Biol. Chem. 28, 3–10.

Chen, J., Delongchamp, R., Tsai, C.-A., Hsueh, H.-M., Sistare, F., Thompson, K.L., Desai, V.G., Fuscoe, J.C., 2004. Analysis of variance components in gene expression data. Bioinformatics 20, 1436–1446.

Chiappetta, P., Roubaud, M.C., Torrésani, B., 2004. Blind source separation and the analysis of microarray data. J. Comput. Biol. 11, 1090–1109.

Curtis, R.K., Brand, M.D., 2004. Analysing microarray data using modular regulation analysis. Bioinformatics 20, 1272–1284.

Datta, S., Datta, S., 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19, 459–466.

Davis, S.J., Millar, A.J., 2001. Watching the hands of the *Arabidopsis* biological clock. Genome Biol. 2, e-pub.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., Moreau, Y., 2002. Adaptive quality-based clustering of gene expression profiles. Bioinformatics 18, 735–746.

Deutsch, J.M., 2003. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics 19, 45–52.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. PNAS 95, 14863–14868.

Ekins, R.P., 1989. Multi-analyse immunoassay. J. Pharm. Biomed. Anal. 7, 155–168.

Ekins, R.P., Chu, F., Biggart, E., 1990. Multispot, multianalyte, immunoassay. Ann. Biol. Clin. (Paris) 48, 655–666.

Ekins, R.P., Chu, F.W., 1991. Multianalyse microspot immunoassay—microanalytical "compact disk" of the future. Clin. Chem. 37, 1955–1967.

Fisher, R.A., 1951. The Design of Experiments, 6th ed. Oliver and Boyd, London.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Heyer, L.J., Kruglyak, S., Yooseph, S., 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9, 1106–1115.

Hirai, M.Y., Yano, M., Goodenowe, D., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., Saito, K., 2004. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. PNAS 101, 10205–10210.

Hoyle, D., Rattray, M., Jupp, R., Brass, A., 2002. Making sense of microarray data distributions. Bioinformatics 18, 576–584.

Hwang, D., Schmitt, W., Stephanopoulos, G., Stephanopoulos, G., 2002. Determination of minimum sample size and discriminatory expression patterns in microarray data. Bioinformatics 18, 1184–1193.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Networks 13, 411–430.

Jarvis, J., Dozmorov, I., Jiang, K., Frank, M.B., Szodoray, P., Alex, P., Centola, M., 2004. Novel approaches to gene expression analysis of active polyarticular juvenile rheumatoid arthritis. Arthritis Res. Ther. 6, R15–R32.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Sha, Q., Chen, J., Tsai, C.-J., Zhang, S., 2004. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 5, 81, e-pub.

Joos, L., Eryuksel, E., Brutsche, M.H., 2003. Functional genomics and gene microarrays—the use in research and clinical medicine. Swiss Med. Wkly. 133, 31–38.

Jordan, B., 2002. Historical background and anticipated developments. Ann. NY Acad. Sci. 975, 24–32.

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P., Castrén, E., 2003. Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics 4, 48, e-pub.

Kendall, M., Stuart, A., Ord, J.K., 1983. The Advanced Theory of Statistics, vol. 3: Design and Analysis, and Time-series. Charles Griffin & Co.

Kerr, K., Churchill, G., 2001. Statistical design and the analysis of gene expression microarray data. Genet. Res. 77, 123–128.

Kerr, K., Martin, M., Churchill, G., 2000. Analysis of variance for gene expression microarray data. J. Comput. Biol. 7, 819–837.

Kim, K.-Y., Kim, B.-J., Yi, G.-S., 2004. Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics 5, 160, e-pub.

Konishi, T., 2004. Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. BMC Bioinformatics 5, 5, e-pub.

Kutalik, Z., Inwald, J., Gordon, S.V., Hewinson, R.G., Bucher, P., Hinds, J., Cho, K.-H., Wokenhauer, O., 2004. Advanced significance analysis of microarray data based on weighted resampling: a comparative study and application to gene deletions in *Mycobacterium bovis*. Bioinformatics 20, 357–363.

Lawrence, N., Milo, M., Niranjan, M., Rashbass, P., Soullier, S., 2004. Reducing the variability in cDNA microarray image processing by Bayesian inference. Bioinformatics 20, 518–526.

Lee, H., Hsu, A., Sajdak, J., Qin, J., Pavlidis, P., 2004. Coexpression analysis of human genes across many microarray data sets. Genome Res. 14, 1085–1094.

Liebermeister, W., 2002. Linear modes of gene expression determined by independent component analysis. Bioinformatics 18, 51–60.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., Lockhart, D.J., 1999. High density synthetic oligonucleotide arrays. Nat. Genet. 21, 20–24.

Lyng, J., Badiee, A., Svendsrud, D., Hovig, E., Myklebost, O., Stokke, T., 2004. Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction. BMC Genomics 5, 10, e-pub.

Martoglio, A.-M., Miskin, J., Smith, S., MacKay, D., 2002. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. Bioinformatics 18, 1617–1624.

Mather, K., 1943. Statistical Analysis in Biology, 1st ed. Methuen.

McCune, H.J., Donaldson, A.D., 2003. DNA replication: telling time with microarrays. Genome Biol. 4, 204, e-pub.

Neuhäuser, M., Senske, R., 2004. The Baumgartner–Weiß–Schindler test for the detection of differentially expressed genes in replicated microarray experiments. Bioinformatics 20, 3553–3564.

Oba, S., Sato, M.-A., Takemasa, I., Monden, M., Matsubara, K.-I., Ishii, S., 2003. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19, 2088–2096.

Oshima, T., Wade, C., Kawagoe, Y., Ara, T., Maeda, M., Masuda, Y., Hiraga, S., Mori, H., 2002. Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in *Escherichia coli*. Mol. Microbiol. 45, 673–695.

Ouyang, M., Welsh, W.J., Georgopoulos, P., 2004. Gaussian mixture clustering and imputation of microarray data. Bioinformatics 20, 917–923.

Park, P.J., Butt, A.J., Kohane, I.S., 2002. Comparing expression profiles of genes with similar promoter regions. Bioinformatics 18, 1576–1584.

Querec, T., Stoyanova, R., Ross, E., Patriotis, C., 2004. A novel approach for increasing sensitivity and correcting saturation artifacts of radioactively labeled cDNA arrays. Bioinformatics 20, 1955–1961.

Renn, S., Aubin-Horth, N., Hofmann, H., 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. BMC Genomics 5, 42, e-pub.

Riva, A., Delorme, M.-O., Chevalier, T., Guilhot, N., Henaut, C., Henaut, A., 2004. The difficult interpretation of transcriptome data: the case of the GATC regulatory network. Comput. Biol. Chem. 28, 109–118.

Sasik, R., Calvo, E., Corbeil, J., 2002. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. Bioinformatics 18, 1633–1640.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Sekowska, A., Robin, S., Daudin, J.J., Henaut, A., Danchin, A., 2001. Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. Genome Biol. 2, Research 0019, e-pub.

Shen, R., Ghosh, D., Chinnaiyan, A.M., 2004. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics 5, 94.

Somorjai, R.L., Dolenko, B., Baumgartner, R., 2003. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 19, 1484–1491.

Sontag, E., Kiyatkin, A., Kholodenko, B., 2004. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. Bioinformatics 20, 1877–1886.

Southern, E.M., Maskos, U., Elder, J.K., 1992. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics 13, 1008–1017.

Stanewsky, R., 2003. Genetic analysis of the circadian system in *Drosophila melanogaster* and mammals. J. Neurobiol. 54, 111–147.

Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21, 631–643.

Stoyanova, R., Querec, T., Brown, T., Patriotis, C., 2004. Normalization of single-channel DNA array data by principal component analysis. Bioinformatics 20, 1772–1784.

Templin, M.F., Stoll, D., Schrenk, M., Traub, P.C., Vohringer, C.F., Joos, T.O., 2002. Protein microarray technology. Trends Biotechnol. 20, 160–166.

Thygesen, H., Zwinderman, A., 2004. Comparing transformation methods for DNA microarray data. BMC Bioinformatics 5, 77, e-pub.

Turk, R., t'Hoen, P.A., Sterrenburg, E., de Menezes, R.X., de Meijer, E.J., Boer, J.M., van Ommen, G.-J.B., den Dunnen, J.T., 2004. Gene expression variation between mouse inbred strains. BMC Genomics 5, 57, e-pub.

Tusher, V., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98, 5116–5121.

Vingron, M., 2001. Bioinformatics needs to adopt statistical thinking. Bioinformatics 17, 389–390.

Vrana, K.E., Freeman, W.M., Aschner, M., 2003. Use of microarray technologies in toxicology research. Neurotoxicology 24, 321–332.

Whitehead, A., Crawford, D.L., 2005. Variation in tissue-specific gene expression among natural populations. Genome Biol. 6, R13, e-pub.

Yeung, K.Y., Medvedovic, M., Bumgarner, R., 2004. From co-expression to co-regulation: how many microarray experiments do we need? Genome Biol. 5, R48, e-pub.

Zar, J.H., 1998. Biostatistical Analysis, 4th ed. Pearson Education.

Zhao, Y., Li, M.-C., Simon, R., 2005. An adaptive method for cDNA microarray normalization. BMC Bioinformatics 6, 28, e-pub.

Zhou, X., Wang, X., Dougherty, E.R., 2003. Missing-value estimation using linear and non-linear regression with Bayesian gene selection. Bioinformatics 19, 2302–2307.

Pour conclure cette partie, il ne faut pas perdre de vue que chaque méthode d'analyse adopte un point de vue particulier sur le nuage. Les différentes méthodes présenteront donc, généralement, des résultats différents mais complémentaires. Le Tableau 1, tiré d'une présentation réalisée à l'ASMDA (Applied stochastic models and data analysis) de Brest et détaillée dans le chapitre suivant illustre bien ce propos. Il présente les différents opérons détectés par des méthodes sur une expérience de transcriptome sur *B. subtilis*. L'étude porte sur les gènes activés par la mise en présence de différentes sources de soufre, le méthyl-thioribose ou la méthionine. Si quatre opérons sont détectés par l'ensemble des méthodes utilisées, trois ne sont pas détectés par une des méthodes, trois sont détectés par deux méthodes et six sont spécifiques d'une méthode d'analyse.

Il peut s'avérer intéressant d'employer différentes méthodes d'analyse des données afin d'aborder une vue un peu plus globale des résultats.

Tableau 1 : Comparaison de résultats obtenus par différentes méthodes

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
yqiXYZ	3	1	1	4	3	6
argCJBD carAB argF	7	15	28	29	201	56
argGH ytzD	3	1	1	6	6	2
ahpCF	2	46	7	85	11	13
lctEP	2	26			36	8
levDEFG sacC	5	316	220	287		
sunAT yolIJK	5		634			13
ydcPQRST yddABCDEFGHIJ	15				1313	116
ytmIJKLM hisP ytmO ytnIJ ribR hipO ytnM	12			45	92	
flgM yvyG flgKL yviEF csrA hag	8					509
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19					350
yxbBA yxnB asnH yxaM	5				15	
yvrPONM	4			494		
ycbCD	2			40		
comGABCDEFG yqzE	8			49		
Relevant detected operons		6	6	9	7	9

Table 2. Comparison of the statistical tools when the experimental factor is identified and fully controlled

2 Le choix de la méthode adéquate pour identifier des gènes différentiellement exprimés : un critère biologique

L'engouement pour le transcriptome a conduit à un développement très important de méthodes spécifiques ou non du transcriptome. Chaque année, plusieurs centaines de publications portent sur l'adaptation ou la création de nouvelles méthodes d'analyse (cf. http://www.nslij-genetics.org/microarray/). Face à cette masse et cette diversité de méthodologie, les biologistes se retrouvent souvent démunis et ont des difficultés à choisir la méthode la plus adéquate pour traiter leurs données. En effet, aucune méthode n'a obtenu jusqu'à présent le consensus général et il n'existe donc pas de protocole idéal d'analyse. La nécessité de comparer l'efficacité des différentes méthodologies disponibles paraît évidente. Différents critères de comparaison ont déjà été appliqués dans le cadre de l'étude du transcriptome.

Dans l'article suivant, nous avons décrit un protocole simple et efficace afin de comparer la fiabilité des résultats obtenus par les méthodes d'analyse des puces à ADN. Ce critère repose sur la structure en opéron du génome bactérien. Les gènes d'un même opéron présentent le même profil d'expression puisqu'ils sont généralement transcrits sur le même ARNm. Les résultats sont identifiés comme cohérents si, lorsque l'on détecte un même opéron comme différentiellement exprimé, les gènes qui le composent ont des rangs de détections proches les uns des autres. La sensibilité et la précision des méthodes d'analyse sont évaluées à partir de ce critère de rang.

Nous avons comparé quatre méthodes sur des données de *Bacillus subtilis* : l'analyse de variance (ANOVA), l'analyse en composantes principales (ACP), l'analyse en composantes indépendantes (ACI) et la régression des moindres carrés partiels (PLS). L'ACI présente les résultats les meilleurs parmi les méthodes testées du point de vue de la sensibilité et la précision. Le protocole a été utilisé afin de tester des méthodes utilisées pour la détection de gènes différentiellement exprimés. Il peut également être appliqué, sans modification, à d'autres types d'analyses comme la co-expression des gènes.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry 28 (2004) 3–10
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA

Anne-Sophie Carpentier*, Alessandra Riva, Pierre Tisseur, Gilles Didier, Alain Hénaut

Laboratoire Génome et Informatique, UMR 8116, Tour Evry2, 523 Place des Terrasses, 91034 Evry, France

Received 5 September 2003; received in revised form 3 December 2003; accepted 3 December 2003

Abstract

The number of statistical tools used to analyze transcriptome data is continuously increasing and no one, definitive method has so far emerged. There is a need for comparison and a number of different approaches has been taken to evaluate the effectiveness of the different statistical tools available for microarray analyses.

In this paper, we describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. It exploits the fact that genes within an operon exhibit the same expression patterns. In order to compare the tools, the genes are ranked according to the most relevant criterion for each tool; for each tool we look at the number of different operons represented within the first twenty genes detected. We then look at the size of the interval within which we find the most significant genes belonging to each operon in question. This allows us to define and estimate the sensitivity and accuracy of each statistical tool.

We have compared four statistical tools using *Bacillus subtilis* expression data: the analysis of variance (ANOVA), the principal component analysis (PCA), the independent component analysis (ICA) and the partial least square regression (PLS). Our results show ICA to be the most sensitive and accurate of the tools tested.

In this article, we have used the protocol to compare statistical tools applied to the analysis of differential gene expression. However, it can also be applied without modification to compare the statistical tools developed for other types of transcriptome analyses, like the study of gene co-expression.

Keywords: Operon; Criterion of comparison; Transcriptome; Expression analysis; ANOVA; ICA; PCA; PLS

1. Introduction

1.1. A word about microarrays

1.1.1. Definition of microarrays

A microarray consist of a solid support on which a series of DNA segments is arranged and fixed in a regular pattern. These segments are incubated with a labeled nucleic acid sample. When a nucleic acid sequence in the sample is complementary to a DNA segment present on the support, it will bind and hybridize to this, specific segment. This hybridization is recorded and analyzed.

1.1.2. The historical background

As Jordan (2002) points out, DNA arrays were already being used in the seventies, in the form of dot blots and slot blots. Ekins et al. developed microspot fluorescent immunoassays in the late eighties and early nineties, proving that the sensitivity of these miniaturized assays was comparable to that of "macroscopic" ones and introducing the concept of micro-array (Ekins, 1989; Ekins et al., 1990; Ekins and Chu, 1991). The concept of miniaturization was also applied to DNA arrays, using two different approaches. One was to deposit the DNA (or complementary DNA) on glass plates, leading to the first publication of a gene expression microarray article in 1995 (Schena et al., 1995). The second approach was that of the oligonucleotide array, where the DNA is directly synthesized onto the support (Fodor et al., 1991; Southern et al., 1992).

* Corresponding author. Tel.: +33-1-60-87-38-74; fax: +33-1-60-87-38-97.
E-mail address: carpentier@genopole.cnrs.fr (A.-S. Carpentier).
URL: http://195.221.65.10:1234/~carpenti/.

1476-9271/$ – see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2003.12.001

1.1.3. Today's microarrays

In the following, "probe" denotes the immobilized DNA on the support and "target" the mobile DNA, cDNA or mRNA. Some authors, however, use the terms the other way round.

The *supports* used for microarrays today are either glass (microscope) slides, (nylon) membranes or silicon chips. The *material fixed* on the support ("probe") can be:

- DNA, representing coding sequences or, more generally, pieces of genomic DNA.
- complementary DNA, obtained from the mRNA of specific genes or expressed sequence tags (ESTs). The latter is usually used for organisms not yet completely sequenced.
- Oligonucleotides; in the case of oligonucleotide arrays the oligos are synthesized directly onto a silicon chip; this process has been pioneered by Affymetrix (see Lipshutz et al. (1999) for a comprehensive review on oligonucleotide arrays).

The *mobile* "*target*" can be:

- DNA,
- complementary DNA (cDNA), obtained from mRNA by reverse transcriptase-PCR (RT-PCR),
- mRNA; this can be used although cDNA is generally preferred.

A hybridization mixture is obtained by labeling the target fluorescently or radioactively. This mixture is then incubated with the prepared microarray and allowed to hybridize with the probe. Finally, the resulting signal intensity, that correlates with the amount of captured probe, is measured, stored in a computer and then analyzed.

Recently, efforts have been made to extend the microarray technology to the field of proteins. The interested reader may refer to the review written by Templin et al. (2002) for a comprehensive introduction to this field.

1.2. Applications

Microarrays can be used for the detection of mutations, DNA sequencing and the analysis of gene expression. The latter application has been gaining in importance and we will focus our attention on this aspect. As microarrays allow measuring the expression levels of thousands of genes at the same time, this opens the possibility to identify differentially expressed genes (Callow et al., 2000) and to cluster those genes sharing similar expression patterns (Heyer et al., 1999). They have become a widespread tool for analyzing the relative transcription levels of genes.

Microarrays have a widespread use, including:

- clinical medicine (see Joos et al. (2003) for a review on this subject);
- the study of the cell-cycle (see for example McCune and Donaldson, 2003);
- the study of the circadian rhythm in animals (see for example Stanewsky, 2003) and plants (see for example Davis and Millar, 2001); and
- the study of plant metabolism (see for example Buckhout and Thimm, 2003).

For further information on microarray technology, the reader may refer to recent review articles (Barrett and Kawasaki, 2003; Vrana et al., 2003); he may also refer to a related NCBI web page (http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html).

1.3. The analysis of the microarray data

Different tools have been developed for or adapted to the analysis of the huge amount of data created in microarray experiments (Draghici, 2002). The number of tools is continuously increasing and no one, definitive method has so far emerged, as is exemplified by the web-site maintained by Li, which has a continuously growing collection of articles on microarray data analysis (http://www.nslij-genetics.org/microarray/).

There is a need of comparing the tools, but identifying an unbiased and biologically relevant criterion for the comparison is difficult (He et al., 2003). A number of different approaches has been taken to compare the effectiveness, or reliability, of the different statistical tools available for microarray analyses.

Some are based on artificial data to define precisely the specificity and sensitivity of these statistical tools (Pan, 2003; Reiner et al., 2003).

Others are based on experimental data. The quality of a statistical tool can be measured by the number of differentially expressed genes which it reveals. A statistical parameter like the *P*-value may be used (Pan, 2002).

Finally some authors combine two criteria, the number of identified genes and their physiological coherence, based on an a priori knowledge of the biological phenomenon studied (Troyanskaya et al., 2002).

1.4. This paper

In this paper, we try to establish a protocol for the comparison of statistical tools (available for microarray analysis) which is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It is based on the expression coherence of genes belonging to the same operon. In bacteria, a number of genes are organized in operons, that is to say clusters of contiguous genes transcribed from one promoter.

A good and reliable statistical tool is one that, when detecting an over- or under-expression for a gene belonging to an operon, also detects this pattern for the other genes belonging to this operon. Indeed, it has been shown that the genes within an operon exhibit the same expression patterns (Murray et al., 2001; Sabatti et al., 2002; Wei et al., 2001; Zimmer et al., 2000).

This criterion, based on the expression coherence of genes belonging to the same operon, therefore reflects a biological property that is not bound to a particular set of experimental conditions. Furthermore, it is independent of the statistical laws (for example Gaussian) governing the variations of the gene expression.

We have tested this criterion on four statistical tools using *Bacillus subtilis* expression data (Sekowska et al., 2001): The analysis of variance (ANOVA), the principal component analysis (PCA), the independent component analysis (ICA) and the partial least square regression (PLS). Note: ANOVA and PLS need the a priori definition of factors, which could influence the level of gene expression; ICA and PCA do not need the definition of any factor for their use.

Two of these tools (ANOVA and PCA) are frequently used for microarray analyses. The other two methods tested (ICA and PLS) have only been recently applied to the analysis of microarray data. All of these methods are used in many other fields.

- The analysis of variance is a classical statistical method for the analysis of fully crossed factorial designs. Its use on microarray data has allowed the identification of differentially expressed genes (Kerr and Churchill, 2001; Kerr et al., 2000).
- The principal component analysis is used to reduce gene space dimension and allows the detection of the major sources of variation (Landgrebe et al., 2002; Peterson, 2003).
- Originally developed for chemometric data (Wold, 1973), the term partial least square regression regroups several methods. PLS has been used in proteome and transcriptome analysis to classify benign and malignant tumours (Alaiya et al., 2000; Cho et al., 2002; Musumarra et al., 2001) or to reduce gene space (Nguyen and Rocke, 2002). In this article, we use PLS to identify differentially expressed genes.
- Independent component analysis (ICA) was originally developed (Comon, 1994) for analyses related to the "cocktail party problem". Its applications in transcriptome analysis include the identification of groups of genes implicated in cancer, the study of the cell cycle (Liebermeister, 2002) and to identify genes that are potentially co-regulated (Chiappetta et al., 2002(personal communication), http://www.cmi.univ-mrs.fr/~torresan/publi.html).

In this article, we set out to compare the four statistical tools mentioned above. However, our method of comparison may be applied to any other statistical tool used in the analysis of microarray data.

2. Methods

2.1. Data

The microarray data used in this study stem from experiments on the sulphur metabolism of *B. subtilis* (Sekowska et al., 2001). The experiments were carried out using Panorama nylon filters *B. subtilis* gene arrays (Sigma–GenoSys Biotechnologies); each array contained all of *B. subtilis'* genes and one gene is represented by one spot. Each gene spot is represented twice on the array.

The aim of these experiments was to identify the genes differentially expressed when the bacteria are grown with methionine or methyl-thioribose as sulphur source. The experiments followed a fully crossed factorial design (Fig. 1) with four factors (sulphur source, day of experiment, amount of RNA used and duplicate of each spot). The data (raw levels of expression) were gathered in an array of 4107 rows (all *B. subtilis* genes) and 16 columns (experimental conditions).

We have used the logarithm (base 10) of these raw data in order to remove much of the proportional relationship between random error and signal intensity (Nadon and Shoemaker, 2002). We have normalized the data (mean equal to 0 and variance equal to 1 for each experimental condition) because two methods (PCA and ICA) need normalized data.

In some parts of the article, the data will be referred to as a cloud of 4107 points (the genes) in a 16-dimensional space (the experimental conditions). In this paper, we will not exploit the dual representation (the 16 experiments in the 4107-dimensional space).

2.2. Programs used

ANOVA, PLS and PCA were carried out using a program called GeneANOVA (Didier et al., 2002). ICA is an adaptation of FASTICA Hyvarinen's fixed-point algorithm (Hyvarinen, 1999) made by Chiappetta and Torrésani (Chiappetta et al., 2002 (personal communication), http://www.cmi.univ-mrs.fr/~torresan/publi.html).

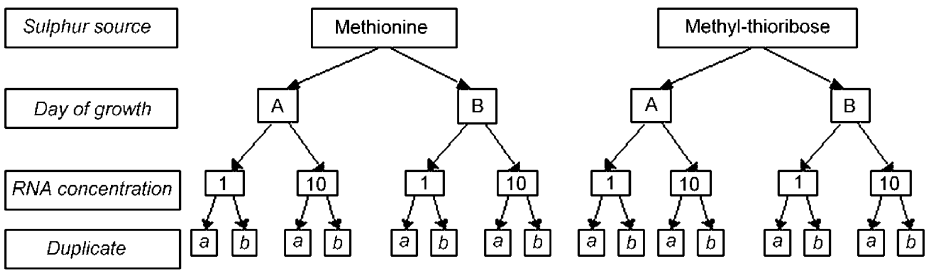


Fig. 1. Experimental design of the transcriptome analysis on *Bacillus subtilis* (Sekowska et al., 2001).

2.3. Choice of parameters

We have chosen to analyze the expression data for the two experimental factors "sulphur source" and "day of experiment".

For ICA and PCA, the axes which correspond to these two factors are determined a posteriori: one determines the relative weight of each of the 16 components whose combination defines the axes; the axes retained are those where either the component "sulphur source", or the component "day of experiment" plays a major role. The factor "day" corresponds to the third axis and the factor "sulphur source" to the fifth. The fourth axis corresponds to an interaction between these two factors.

For each gene, the equation used for ANOVA is the following:

$$Y_{ijkl} = \mu + S_i + J_j + C_k + D_l + \varepsilon_{ijkl}$$

where Y_{ijkl} is the gene intensity; μ the mean of the intensities of expression measured for the gene; S_i, J_j, C_k and D_l are, respectively, the effects of sulphur source i, experiment day j, RNA concentration k and duplicate l on the gene intensity; and ε_{ijkl} is the residual error.

There are 16 measurements per gene. Five degrees of freedom are lost for the estimation of the mean and the variances of the four factors. The residual variance ε_{ijkl} has 11 degrees of freedom. It encompasses all interactions: between two factors (6), between three factors (4) and between four factors (1).

F = "variance of the factor of interest"/"residual variance" with one degree of freedom in the numerator and 11 degrees of freedom in the denominator.

The interactions between the factors were not estimated because of the experimental design and the low degree of freedom obtained.

2.4. Operons

We need to know how the genes of *B. subtilis* are organized into operons. A presumed operon is defined as a group of contiguous genes that are on the same reading strand delimited either by a promoter and a terminator (predicted or not) or a gene, which lies on the other DNA strand. This allowed to find the operons in *B. subtilis* (Subtilist, http://genolist.pasteur.fr/SubtiList/). We compiled a list in which each gene is either assigned to an operon or defined as an isolated gene. This list may be consulted at http://195.221.65.10:1234/∼carpenti/. Even if some predicted operons will prove to be artefacts, this will only introduce a systematic bias for all the statistical tools tested. This will not raise any problem for the comparison of the statistical tools and it will not influence our conclusions about the quality of these tools with respect to each other.

2.5. Evaluating procedure

To compare statistical tools, one needs to define quantitative criteria that will measure the "tool reliability": sensitivity, accuracy and the detection of false positives need to be evaluated.The following procedure was applied:

1. The genes are ranked as a function of their expression changes (rank #1 is the most significant).
2. "Detected Operons" are identified based on the ranks (one gene with rank ≤ 20 and another gene with rank ≤ 100).
3. The most significant interval (MSI) is determined.
4. False positives are evaluated (MSI ≥ 700).
5. "Relevant detected operons" are identified (MSI < 700).
6. The accuracy of a "relevant detected operon" is evaluated (MSI < 150).
7. The sensitivity of a tool is evaluated.

2.5.1. Ranking of the genes

In order to compare the four tools under the best possible conditions, the genes are ranked according to the most relevant criterion for each tool, that is to say, the one that gives the most coherent results for the tool:

- for ANOVA, the *P*-value obtained for each gene;
- for PLS, the weight of the gene for the axis determination; and
- for PCA and ICA, the remoteness from the cloud center of the projection of the gene on the axis studied.

We thus obtain for each tool a list of genes, ranked according to a specific criterion; the most significant gene has rank #1. The order of the genes on the lists obtained may differ from each other.

2.5.2. Identification of "detected operons"

We define an operon to be detected ("detected operon") by a tool if at least one of its genes has a rank $\leq$20 and another of its genes a rank $\leq$100. For the assignment of genes to operons, we have used the list which may be consulted at http://195.221.65.10:1234/∼carpenti/. It should be noted that a priori the "detected operons" may be different for the various tools tested.

A possible bias of this method presents itself in one particular case: If one of the "detected operons" is very large, a considerable proportion of the genes with a rank $\leq$20 will belong to this particular operon, leaving "no place" for the other operons to be detected. The same problem may arise if a large number of isolated genes (not belonging to an operon) are highly relevant. As this possible bias will be present for all four statistical tools tested, it will not raise any problem for the comparison of the statistical tools and it will not influence our conclusions about the quality of these tools with respect to each other.

Note: the choice of "20; 100" is an arbitrary one. In order to establish whether this choice might affect the results and thus the conclusions of this paper, we have also run through

the procedure using, successively "10; 50" and "40; 200" for the identification of "detected operons". The results may be consulted at http://195.221.65.10:1234/~carpenti/ (see also Section 3).

2.5.3. Determination of the most significant interval

In order to facilitate the analysis and comparison of the statistical tools, we introduce the most significant interval (MSI). It is calculated for each "detected operon" in the following manner:

$$\mathrm{MSI}_j = \mathrm{median}_j - \mathrm{first}_j$$

where MSI_j is the MSI of "detected operon" j, median_j is the median of the rank values of the genes belonging to "detected operon" j, and first_j is the smallest rank value within "detected operon" j

2.5.4. Evaluation of false positives

The reliability of a statistical tool will also be measured by the absence of false positives.

For the definition of false positives, we exploit the fact that each gene spot had been duplicated on the microarrays and any difference measured for two spots belonging to the same gene cannot have a biological cause. We ranked the genes according to this "duplicate factor '', as described under point 1 and identified "detected operons" as described under point 2. As there is no biological cause for this detection, we find ourselves with false positives.

As before, a priori the false positives detected may be different for the various tools tested.

The results of this analysis lead us to conclude that a "detected operon" is a false positive when MSI $\geq$ 700 (see Table 1 for details).

Table 1
Quantification of false positives

Operon name	Operon size	MSI (most significant interval)			
		ANOVA	PLS	PCA	ICA
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19	2385	2243	1193	2499
yonRSTUVX yopAB	8	61	134	127	251
hemAXCDBL	6	1360	1547		
ruvAB queA tgt yrbF	5			1005	707

For the definition of false positives we exploit the fact that each gene spot had been duplicated on the microarrays and any difference measured for two spots belonging to the same gene cannot have a biological cause. We ranked the genes according to this "duplicate factor", as a function of the differences in their expressions, then identified "detected operons" and calculated the MSI (see Section 2 for details). As there is no biological cause for this detection, we find ourselves with false positives; they are characterized by a large MSI; this leads us to conclude that a "detected operon" is a false positive when MSI $\geq$ 700. One exception is the operon *YonRSTUVXyopAB*, detected by all four tools, with small MSIs. As we cannot give a biological reason, we suspect that its detection is due to a default on the microarray used in the experiments.

2.5.5. Identification of "relevant detected operons"

The definition of "relevant detected operons" follows from the definition of false positives:

"relevant detected operons" have an MSI $<$ 700.

2.5.6. Evaluating the accuracy of a "relevant detected operon"

We define that an operon is detected with good accuracy if its MSI is lower then a given threshold. This threshold was determined such that 80% of the "detected operons" have a MSI below the threshold. Our results lead us to state that: Operons detected with good accuracy have an MSI $<$ 150.

2.5.7. Evaluating the sensitivity of the tools

The sensitivity of the tools is estimated by comparing the number or "relevant detected operons" identified by each tool.

2.6. The comparison of the tools under three typical experimental conditions

We have decided to compare the four statistical tools under three experimental conditions biologists are frequently faced with:

- *The experimental factor is identified and fully controlled*: In the case of the microarray data used in this study, this factor is the sulphur source contained in the growth medium. In one case the sulphur source was methionine, in the other case it was methylthioribose. These two compounds are metabolically closely related. The four statistical tools were tested on these experimental data. The results obtained are displayed in Table 2.
- *The experimental factor is identified but not under control*: In this case it was "day". The experiments were carried out twice, on different days. The protocol followed was the same on these 2 days; however, parameters like "room temperature" were not necessarily the same, thus introducing a factor in the experimental setup that was identified but not under control. The results obtained are displayed in Table 3.
- *The interaction between experimental factors*: The aim of a protocol is to separate completely the different experimental factors. However, the expression of certain genes may be under the control of more than one factor. In this case, one talks of an "interaction between experimental factors". ANOVA and PLS are adapted to the analysis of variations due to a single experimental factor; they are not well suited for the study of interactions between factors; they were not tested under this condition. On the other hand, ICA and PCA are well adapted to cope with possible interactions; these interactions are identified because more than one factor plays a major role in the definition of an axis. The results obtained are displayed in Table 4.

Table 2
Comparison of the statistical tools when the experimental factor is identified and fully controlled

Operon name	Operon size	MSI (most significant interval)			
		ANOVA	PLS	PCA	ICA
yqiXYZ	3	1	1	3	6
argCJBD carAB argF	7	15	28	201	56
argGH ytzD	3	1	1	6	2
ahpCF	2	46	7	11	13
lctEP	2	26		36	8
levDEFG sacC	5	316	220		
sunAT yolIJK	5		**635**		13
ydcPQRST yddABCDEFGHIJ	15			**1313**	116
flgM yvyG flgK yviE yviF csrA hag	8				509
yxbBA yxnB asnH yxaM	5			15	
ytmIJKLM hisP ytmO ytnIJ ribR hipO ytnM	12			92	
fliLMY cheYfliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19				350
Relevant detected operons		6	6	7	9

The identified and controlled experimental factor is the sulphur source (either methionine, or methylthioribose). Genes were ranked as a function of the differences in their expressions, false positives (MSI $\geq$ 700) and "relevant detected operons"(MSI $<$ 700) were identified (see Section 2 for details). The bold entry for PLS, with MSI = 635 is estimated to be a borderline case for a false positive; it has been included for PLS's total of "relevant detected operons". Note that only PCA detects a false positive (shaded entry). ICA is the most sensitive tool under these experimental conditions, identifying the largest number of "relevant detected operons". ANOVA and PLS are the least sensitive.

3. Results and discussion

Microarrays are defined as a tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. They are widely used for analyzing the relative transcription level of genes. The number of statistical tools for analyzing the huge amount of data created in the experiments is continuously growing and no-one of these tools has yet emerged as the definitive one.

We have developed a protocol for the comparison of statistical tools applied to the analysis of transcription data. We have applied this method to compare four statistical tools (ANOVA, PLS, ICA and PCA) under three typical experimental conditions. All four tools were compared under two of these conditions (see Tables 2 and 3 for details), whilst only ICA and PCA, which do not need the a priori definition of experimental factors, could be tested under the third condition (see Table 4 for details).

Based on our observations, we have defined threshold values to define "relevant detected operons" (MSI $<$ 700), false positives (MSI $\geq$ 700) and to define a good accuracy (MSI $<$ 150); the sensitivity of the tools is estimated by comparing the number of "relevant detected operons" identified by each tool.

Table 3
Comparison of the statistical tools when the experimental factor is identified but not under control

Operon name	Operon size	MSI (most significant interval)			
		ANOVA	PLS	PCA	ICA
comGABCDEFG yqzE	8	16	26	6	4
comFABC yvyF	4	339		66	19
cotVWXYZ	5		147	315	417
groESL	2		15		
yvaVWXY	4			53	
yqxM sipW cotN	3			79	
comEABC	3				35
Relevant detected operons		2	3	5	4

The experiments were carried out twice, on different days, using the same protocol; however, parameters like "room temperature" were not necessarily the same on the 2 days, introducing an identified but not controlled factor. PCA and ICA are the most sensitive tools, whilst ANOVA is the least sensitive (please refer to the legend of Table 2 for details about the classification procedure).

Table 4
Comparison of the statistical tools to detect possible interactions between the experimental factors

Operon name	Operon size	MSI (most significant interval)	
		PCA	ICA
purMNHD	4	71	57
ybaC rpsJ rplCDWB rpsS rplV rpsC rplP rpmC rpsQ rplNXE rpsNH rplFR rpsE rpmD rplO secY adk map	25	51	56
alsS alsD	2		25
rpsL rpsG fus tufA	4		21
yvaVWXY	4		73
yxbBA yxnB asnH yxaM	5		126
yyaEF rpsF ssb rpsR	5	408	
Relevant detected operons		3	6

The expression of certain genes may be under the control of more than one factor, leading to an interaction between experimental factors. Only ICA and PCA are well adapted to cope with possible interactions; these interactions are identified because more than one factor plays a major role in the definition of an axis. ICA is more sensitive than PCA (please refer to the legend of Table 3 for details about the classification procedure).

Table 5
Overview of the results

	ANOVA	PLS	PCA	ICA
Relevant detected operons				
Tables 2–4	8	9	15	19
Tables 2 and 3	8	9	12	13
Accuracy of detection (%)				
Tables 2–4	75	78	80	84
Tables 2 and 3	75	78	83	77

The table sums up the results obtained in this study. The first part of the table relates to the number of "relevant detected operons" identified and thus to the tools' relative sensitivities. "Tables 2–4": adding the results from Tables 2–4, the total of "relevant detected operons" has been calculated for each tool. The entries for "Tables 2 and 3" have been obtained accordingly. Note that in both cases, ICA has the highest overall sensitivity, identifying the largest number of "relevant detected operons", whilst ANOVA is the least sensitive. The second part of the tables relates to the tools' accuracies: the percentage of "relevant detected operons" identified with a "good accuracy" (MSI < 150) has been calculated for each tool, adding the results from Tables 2–4 ("Tables 2–4"), etc. (see above). Overall, ICA has the highest accuracy, very closely followed by PCA, whilst ANOVA has the lowest accuracy.

Table 5 sums up the results obtained. Overall, we observe that ANOVA has the lowest sensitivity, whilst ICA is the tool with the highest sensitivity. The same observations can be made regarding the accuracies of the tools. It is interesting to note that even under the two experimental conditions for which ANOVA was conceived (Tables 2 and 3), it performs less well than ICA. PLS performs similarly to ANOVA. PCA has an intermediate performance. However, each tool may detect operons not identified by the other tools.

The results obtained by testing the four statistical tools show us that ICA has overall the best performance. This result holds true even if the criteria for "detected operon" are changed (instead of "20; 100" using "10; 50" or "40; 200", results not shown; see http://195.221.65.10:1234/~carpenti/ for details).

In this paper, we have set out to describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. The criterion used in our method is based on the expression coherence of genes belonging to the same operon. The method is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It allows to compare the sensitivity, the accuracy and the detection of false positives of different statistical tools. As it is a comparative method, any bias linked to the criterion (for example uncertainties about the reality of a predicted operon) will influence in the same way the results obtained for each of the tools tested.

Here, we have used this method to compare statistical tools applied to the analysis of differential gene expression. However, the above protocol can also be applied without modification to compare the statistical tools developed for other types of transcriptome analyses, like the study of gene co-expression.

Acknowledgements

We are grateful to Antoine Danchin and Agnieszka Sekowska for having provided us with their data and to Bruno Torrésani and Pierre Chiappetta for the ICA program that they adapted to gene expression analysis. This work was supported by the French Industry Ministry contract ASG number 01 4 90 6093.

References

Alaiya, A.A., Franzen, B., Hagman, A., Silfversward, C., Moberger, B., Linder, S., Auer, G., 2000. Classification of human ovarian tumors using multivariate data analysis of polypeptide expression patterns. Int. J. Cancer 86 (5), 731–736.

Barrett, J.C., Kawasaki, E.S., 2003. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discov. Today 8 (3), 134–141.

Buckhout, T.J., Thimm, O., 2003. Insights into metabolism obtained from microarray analysis. Curr. Opin. Plant Biol. 6 (3), 288–296.

Chiappetta, P., Roubaud, M.C., Torrésani, B., 2002. Blind Source Separation de Sources and the Analysis of Microarray Data, personal communication. http://www.cmi.univ-mrs.fr/~torresan/publi.html.

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., Rubin, E.M., 2000. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res. 10 (12), 2022–2029.

Cho, J.H., Lee, D., Park, J.H., Kim, K., Lee, I.B., 2002. Optimal approach for classification of acute leukemia subtypes based on gene expression data. Biotechnol. Prog. 18 (4), 847–854.

Comon, P., 1994. Independent component analysis—a new concept? Signal Process. 36, 287–314.

Davis, S.J., Millar, A.J., 2001. Watching the hands of the Arabidopsis biological clock. Genome Biol. 2 (3), e-pub.

Didier, G., Brezellec, P., Remy, E., Henaut, A., 2002. GeneANOVA–gene expression analysis of variance. Bioinformatics 18 (3), 490–491.

Draghici, S., 2002. Statistical intelligence: effective analysis of high-density microarray data. Drug Discov. Today 7 (11), S55–S63.

Ekins, R.P., 1989. Multi-analyte immunoassay. J. Pharm. Biomed. Anal. 7 (2), 155–168.

Ekins, R.P., Chu, F., Biggart, E., 1990. Multispot, multianalyte, immunoassay. Ann. Biol. Clin. (Paris) 48 (9), 655–666.

Ekins, R.P., Chu, F.W., 1991. Multianalyte microspot immunoassay—microanalytical "compact disk" of the future. Clin. Chem. 37 (11), 1955–1967.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251 (4995), 767–773.

He, Y.D., Dai, H., Schadt, E.E., Cavet, G., Edwards, S.W., Stepaniants, S.B., Duenwald, S., Kleinhanz, R., Jones, A.R., Shoemaker, D.D., Stoughton, R.B., 2003. Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. Bioinformatics 19 (8), 956–965.

Heyer, L.J., Kruglyak, S., Yooseph, S., 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9 (11), 1106–1115.

Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Networks 10 (3), 626–634.

Joos, L., Eryuksel, E., Brutsche, M.H., 2003. Functional genomics and gene microarrays—the use in research and clinical medicine. Swiss. Med. Wkly. 133 (3–4), 31–38.

Jordan, B., 2002. Historical background and anticipated developments. Ann. NY Acad. Sci. 975, 24–32.

Kerr, M.K., Churchill, G.A., 2001. Statistical design and the analysis of gene expression microarray data. Genet Res. 77 (2), 123–128.

Kerr, M.K., Martin, M., Churchill, G.A., 2000. Analysis of variance for gene expression microarray data. J. Comput. Biol. 7 (6), 819–837.

Landgrebe, J., Wurst, W., Welzl, G., 2002. Permutation-validated principal components analysis of microarray data. Genome Biol. 3 (4), 0019.0011–0019.0011.

Liebermeister, W., 2002. Linear modes of gene expression determined by independent component analysis. Bioinformatics 18 (1), 51–60.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., Lockhart, D.J., 1999. High density synthetic oligonucleotide arrays. Nat Genet. 21 (1 Suppl.), 20–24.

McCune, H.J., Donaldson, A.D., 2003. DNA replication: telling time with microarrays. Genome Biol. 4 (2), 204.

Murray, A.E., Lies, D., Li, G., Nealson, K., Zhou, J., Tiedje, J.M., 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc. Natl. Acad. Sci. U.S.A. 98 (17), 9853–9858.

Musumarra, G., Condorelli, D.F., Scire, S., Costa, A.S., 2001. Shortcuts in genome-scale cancer pharmacology research from multivariate analysis of the National Cancer Institute gene expression database. Biochem. Pharmacol. 62 (5), 547–553.

Nadon, R., Shoemaker, J., 2002. Statistical issues with microarrays: processing and analysis. Trends Genet. 18 (5), 265–271.

Nguyen, D.V., Rocke, D.M., 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18 (1), 39–50.

Pan, W., 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18 (4), 546–554.

Pan, W., 2003. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. Bioinformatics 19 (11), 1333–1340.

Peterson, L.E., 2003. Partitioning large-sample microarray-based gene expression profiles using principal components analysis. Comput. Methods Programs Biomed. 70 (2), 107–119.

Reiner, A., Yekutieli, D., Benjamini, Y., 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19 (3), 368–375.

Sabatti, C., Rohlin, L., Oh, M.K., Liao, J.C., 2002. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res. 30 (13), 2886–2893.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270 (5235), 467–470.

Sekowska, A., Robin, S., Daudin, J.J., Henaut, A., Danchin, A., 2001. Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. Genome Biol. 2 (6), 0019.0011–0019.0012.

Southern, E.M., Maskos, U., Elder, J.K., 1992. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics 13 (4), 1008–1017.

Stanewsky, R., 2003. Genetic analysis of the circadian system in *Drosophila* melanogaster and mammals. J. Neurobiol. 54 (1), 111–147.

Templin, M.F., Stoll, D., Schrenk, M., Traub, P.C., Vohringer, C.F., Joos, T.O., 2002. Protein microarray technology. Trends Biotechnol. 20 (4), 160–166.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D., Altman, R.B., 2002. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18 (11), 1454–1461.

Vrana, K.E., Freeman, W.M., Aschner, M., 2003. Use of microarray technologies in toxicology research. Neurotoxicology 24 (3), 321–332.

Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A., LaRossa, R.A., 2001. High-density microarray-mediated gene expression profiling of Escherichia coli. J. Bacteriol. 183 (2), 545–556.

Wold, H., 1973. Nonlinear iterative partial least squares (NIPALS) modelling—some current development. In: Krishnajah, P.R. (Ed.), Multivariate Analysis, Academic Press, New York, pp. 383–407.

Zimmer, D.P., Soupene, E., Lee, H.L., Wendisch, V.F., Khodursky, A.B., Peter, B.J., Bender, R.A., Kustu, S., 2000. Nitrogen regulatory protein C-controlled genes of Escherichia coli: scavenging as a defense against nitrogen limitation. Proc. Natl. Acad. Sci. U.S.A. 97 (26), 14674–14679.

Ces résultats ont été complétés dans le cadre de l'ASMDA (conférence sur Applied stochastic models and data analysis) de Brest. Outre les méthodes déjà comparées, nous avons rajouté le t-test et le t-test apparié. Il en ressort que l'ANOVA et le t-test donne des résultats comparables et que de toutes les méthodes supervisées, le t-test apparié présente les meilleurs résultats. L'ACI reste néanmoins globalement la méthode la plus performante sur ce jeu de données.

The operons, a criterion to compare the reliability of transcriptome analysis tools

A.-S. Carpentier, A. Riva, G. Didier, J.-L. Risler, and A. Hénaut

Laboratoire Génome et informatique UMR 8116
Tour Evry2, 523 Place des Terrasses
91034 EVRY, France
(e-mail: carpentier@genopole.cnrs.fr)

Abstract. The number of statistical tools used to analyze transcriptome data is continuously increasing and no one, definitive method has so far emerged. There is a need for comparison and a number of different approaches has been taken to evaluate the effectiveness of the different statistical tools available for microarray analyses. In this paper we describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. It exploits the fact that genes within an operon exhibit the same expression patterns. We have compared five statistical tools using *Bacillus subtilis* expression data: ANOVA, PCA, ICA, the *t*-test and the paired *t*-test. Our results show ICA to be the most sensitive and accurate of the tools tested.
Keywords: operon, criterion of comparison, transcriptome, expression analysis.

1 Introduction on microarrays and their analysis

Protein activities are the bases of cell and organism functioning. In order to fit to changes in extern or interne physiological conditions the expression level of some genes and the quantity of the corresponding proteins may vary. As proteins are much harder to analyze than mRNAs, techniques for transcriptome analysis have been more popular up to now. In the last decades a tool has been developed in order to measure the expression levels of many genes (several thousands of genes) at the same time.

As microarrays allow measuring the expression levels of thousands of genes at the same time, this opens the possibility to identify differentially expressed genes [Callow *et al.*, 2000] and to cluster those genes sharing similar expression patterns [Heyer *et al.*, 1999]. This allow the identification of gene functions, regulation and networks.

Different tools have been developed for or adapted to the analysis of the huge amount of data created in microarray experiments. The number of tools is continuously increasing and no one, definitive method has so far emerged. There is a need of comparing the tools, but identifying an unbiased and biologically relevant criterion for the comparison is difficult [He *et al.*, 2003]. A number of different approaches has been taken to compare the effectiveness, or reliability, of the different statistical tools available for microarray analyses:

* Some are based on artificial data to define precisely the specificity and sensitivity of these statistical tools ([Reiner *et al.*, 2003]).

* Others are based on experimental data. The quality of a statistical tool can be measured by the number of differentially expressed genes which it reveals. A statistical parameter like the p-value may be used [Pan, 2002].

* Finally some authors combine two criteria, the number of identified genes and their physiological coherence, based on an a priori knowledge of the biological phenomenon studied [Troyanskaya *et al.*, 2002].

In this paper we try to establish a protocol for the comparison of statistical tools (available for microarray analysis) which is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It is based on the expression coherence of genes belonging to the same operon. In bacteria a number of genes are organized in operons, that is to say clusters of contiguous genes transcribed from one promoter. For an operon a single mRNA corresponds to several genes whereas for isolated genes one mRNA corresponds to one gene. It has been shown that the genes within an operon exhibit the same expression patterns [Sabatti *et al.*, 2002].

That is why, a good and reliable statistical tool is one that, when detecting an over- or under-expression for a gene belonging to an operon, also detects this pattern for the other genes belonging to this operon. This criterion, based on the expression coherence of genes belonging to the same operon, therefore reflects a biological property that is not bound to a particular set of experimental conditions.

We have tested this criterion on five statistical tools using *Bacillus subtilis* expression data [Sekowska *et al.*, 2001]: The Analysis of Variance (ANOVA), the Principal Component Analysis (PCA), the Independent Component Analysis (ICA), the *t*-test and the paired *t*-test. Note: ANOVA and the *t*-tests need the a priori definition of factors, which could influence the level of gene expression; ICA and PCA do not need the definition of any factor for their use.

2 Methods

The microarray data used in this study stem from experiments on the sulphur metabolism of *Bacillus subtilis* [Sekowska *et al.*, 2001]. The experiments were carried out using *B. subtilis* gene arrays; each array contained all of *B. subtilis*' genes and one gene is represented by one spot. Each gene spot is represented twice on the array.

The aim of these experiments was to identify the genes differentially expressed when the bacteria are grown with methionine or methyl-thioribose as sulphur source. The experiments followed a fully crossed factorial design with 4 factors (sulphur source, day of experiment, amount of RNA used and duplicate of each spot).

We have used the logarithm (base 10) of these raw data in order to remove much of the proportional relationship between random error and signal intensity. We have normalized the data (mean equal to 0 and variance equal to 1 for each experimental condition).

We have chosen to analyze the expression data for the two experimental factors "sulphur source" and "day of experiment". For ICA and PCA the axes which correspond to these two factors are determined a posteriori. For PCA the factor "day" corresponds to the third axis and the factor "sulphur source" to the fifth. The fourth axis corresponds to an interaction between these two factors.

For each gene, the model used for ANOVA is the following:

$$Y_{ijkl} = \mu + S_i + J_j + C_k + D_l + \epsilon_{ijkl}$$

where Y_{ijkl} is the gene intensity

μ is the mean of the intensities of expression measured for the gene

S_i, J_j, C_k and D_l are, respectively, the effects of sulphur source i, experiment day j, RNA concentration k and duplicate l on the gene intensity

ϵ_{ijkl} is the residual error.

We need to know how the genes of *Bacillus subtilis* are organized into operons. A presumed operon is defined as a group of contiguous genes that are on the same reading strand delimited either by a promoter and a terminator (predicted or not) or a gene, which lies on the other DNA strand. This allowed to find the operons in *Bacillus subtilis* (Subtilist).

To compare statistical tools, one needs to define quantitative criteria that will measure the "tool reliability": sensitivity, accuracy and the detection of false positives need to be evaluated.

The following procedure was applied:

1. The genes are ranked as a function of their expression changes (rank #1 is the most significant).
   In order to compare the five tools under the best possible conditions, the genes are ranked according to the most relevant criterion for each tool, that is to say, the one that gives the most coherent results for the tool:
   * for ANOVA and the *t*-tests, the p-value obtained for each gene;
   * for PCA and ICA, the remoteness from the cloud centre of the projection of the gene on the axis studied.
   We thus obtain for each tool a list of genes, ranked according to a specific criterion. The order of the genes on the lists obtained may differ from each other.
2. "Detected Operons" are identified based on the ranks (one gene of the operon with rank ≤ 20 and another gene with rank ≤ 100).
   It should be noted that a priori the "Detected Operons" may be different for the various tools tested.
3. The Most Significant Interval (MSI) is determined.

In order to facilitate the analysis and comparison of the statistical tools we introduce the Most Significant Interval (MSI). It is calculated for each "Detected Operon" in the following manner:

$$MSI_j = median_j - first_j$$

Where MSI_j is the MSI of "Detected Operon" j
$median_j$ is the median of the rank values of the genes belonging to "Detected Operon" j
$first_j$ is the smallest rank value within "Detected Operon" j

4. False positives are evaluated (MSI≥700).
   The reliability of a statistical tool will also be measured by the absence of false positives. For the definition of false positives we exploit the fact that each gene spot had been duplicated on the microarrays and any difference measured for two spots belonging to the same gene cannot have a biological cause. We ranked the genes according to this "duplicate factor ", as described under point 1 and identified "Detected Operons" as described under point 2. As there is no biological cause for this detection, we find ourselves with false positives. The results of this analysis lead us to conclude that a "Detected Operon" is a false positive when MSI≥700 (see table 1 for details).

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19	2385	2242	1613	1193	2499
yonRSTUVX yopAB	8	61	134	124	127	251
hemAXCDBL	6	1360	1547			
ruvAB queA tgt yrbF	5				1005	707

Table 1. Quantification of false positives

[We find ourselves with false positives. One exception is the operon yonRSTUVXyopAB, detected by all four tools, with small MSIs. As we cannot give a biological reason, we suspect that its detection is due to a default on the microarray used in the experiments.]

5. "Relevant Detected Operons" are identified (MSI<700). The definition of "Relevant Detected Operons" follows from the definition of false positives: "Relevant Detected Operons" have an MSI<700.
6. The accuracy of a "Relevant Detected Operon" is evaluated (MSI<150). We define that an operon is detected with good accuracy if its MSI is lower then a given threshold. Our results lead us to state that: Operons detected with good accuracy have an MSI<150.
7. The sensitivity of a tool is evaluated.

The sensitivity of the tools is estimated by comparing the number or "Relevant Detected Operons" identified by each tool.

We have decided to compare the five statistical tools under three experimental conditions biologists are frequently faced with:

* The experimental factor is identified and fully controlled. In the case of the microarray data used in this study, this factor is the sulphur source contained in the growth medium. In one case the sulphur source was methionine, in the other case it was methylthioribose. The five statistical tools were tested on these experimental data. The results obtained are displayed in table 2.

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
yqiXYZ	3	1	1	4	3	6
argCJBD carAB argF	7	15	28	29	201	56
argGH ytzD	3	1	1	6	6	2
ahpCF	2	46	7	85	11	13
lctEP	2	26			36	8
levDEFG sacC	5	316	220	287		
sunAT yolIJK	5		634			13
ydcPQRST yddABCDEFGHIJ	15				1313	116
ytmIJKLM hisP ytmO ytnIJ ribR hipO ytnM	12			45	92	
flgM yvyG flgKL yviEF csrA hag	8					509
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19					350
yxbBA yxnB asnH yxaM	5				15	
yvrPONM	4			494		
ycbCD	2			40		
comGABCDEFG yqzE	8			49		
Relevant detected operons		6	6	9	7	9

Table 2. Comparison of the statistical tools when the experimental factor is identified and fully controlled

* The experimental factor is identified but not under control. In this case it was "day". The experiments were carried out twice, on different days. The protocol followed was the same on these two days; however, parameters like "room temperature" were not necessarily the same, thus introducing a factor in the experimental setup that was identified but not under control. The results obtained are displayed in table 3.

* The interaction between experimental factors. The aim of a protocol is to separate completely the different experimental factors. However, the

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
comGABCDEFG yqzE	8	16	26	28	6	4
comFABC yvyF	4	339			66	19
cotVWXYZ	5		148		315	417
groESL	2			37		
yvaVWXY	4				53	
yqxM sipW cotN	3				79	
comEABC	3					35
Relevant detected operons		2	2	2	5	4

Table 3. Comparison of the statistical tools when the experimental factor is identified but not under control

expression of certain genes may be under the control of more than one factor. In this case one talks of an "interaction between experimental factors". ANOVA and the *t*-tests are adapted to the analysis of variations due to a single experimental factor; they are not well suited for the study of interactions between factors; they were not tested under this condition. On the other hand, ICA and PCA are well adapted to cope with possible interactions; these interactions are identified because more than one factor plays a major role in the definition of an axis. The results obtained are displayed in table 4.

Operon name	Operon size	MSI	
		PCA	ICA
purMNHD	4	71	57
ybaC rpsJ rplCDWB rpsS rplV rpsC rplP rpmC rpsQ rplNXE rpsNH rplFR rpsE rpmD rplO secY adk map	25	51	56
alsS alsD	2		25
rpsL rpsG fus tufA	4		21
yvaVWXY	4		73
yxbBA yxnB asnH yxaM	5		126
yyaEF rpsF ssb rpsR	5	408	
Relevant detected operons		3	6

Table 4. Comparison of the statistical tools to detect possible interactions between the experimental factors

3 Results and discussion

Microarrays are defined as a tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. They are widely used for analyzing the relative transcription level of genes. The number of statistical tools for analyzing the huge amount of data created in the experiments is continuously growing and no-one of these tools has yet emerged as the definitive one.

We have developed a protocol for the comparison of statistical tools applied to the analysis of transcription data. We have applied this method to compare five statistical tools (ANOVA, *t*-test, paired *t*-test, ICA and PCA) under three typical experimental conditions. All five tools were compared under two of these conditions (see tables 2 and 3 for details), whilst only ICA and PCA, which do not need the a priori definition of experimental factors, could be tested under the third condition (see table 4 for details).

Based on our observations, we have defined threshold values to define "Relevant Detected Operons" (MSI<700), false positives (MSI≥700) and to define a good accuracy (MSI<150); the sensitivity of the tools is estimated by comparing the number of "Relevant Detected Operons" identified by each tool.

	ANOVA	*t*-test	Paired *t*-test	PCA	ICA
Relevant detected operons					
Table 2-4	8	8	11	15	19
Table 2-3	8	8	11	12	13
Accuracy of Detection					
Table 2-4	75%	75%	82%	80%	84%
Table 2-3	75%	75%	82%	83%	77%

Table 5. Overview of the results

[The table sums up the results obtained in this study. The first part of the table relates to the number of "Relevant Detected Operons" identified and thus to the tools' relative sensitivities. "Tables 2 - 4": adding the results from Tables 2, 3 and 4, the total of "Relevant Detected Operons" has been calculated for each tool. The entries for "Tables 2 - 3" have been obtained accordingly. The second part of the tables relates to the tools' accuracies: the percentage of "Relevant Detected Operons" identified with a "good accuracy" (MSI<150) has been calculated for each tool, adding the results from Tables 2, 3 and 4 ("Tables 2 - 4") etc.]

Table 5 sums up the results obtained. Overall, we observe that ANOVA and *t*-test have the lowest sensitivity, whilst ICA is the tool with the highest sensitivity. The same observations can be made regarding the accuracies of the tools. It is interesting to note that even under the two experimental conditions for which ANOVA and the *t*-test were conceived (tables 2 and 3),

it performs less well than ICA. The paired *t*-test has a high accuracy but a lower sensitivity than ICA just like PCA. However, each tool may detect operons not identified by the other tools.

The results obtained by testing the five statistical tools show us that ICA has overall the best performance.

In this paper we have set out to describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. The criterion used in our method is based on the expression coherence of genes belonging to the same operon. The method is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It allows to compare the sensitivity, the accuracy and the detection of false positives of different statistical tools.

Here we have used this method to compare statistical tools applied to the analysis of differential gene expression. However, the above protocol can also be applied without modification to compare the statistical tools developed for other types of transcriptome analyses, like the study of gene co-expression.

References

[Callow *et al.*, 2000]M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome Res*, pages 2022–9., 2000.

[He *et al.*, 2003]Y. D. He, H. Dai, E. E. Schadt, G. Cavet, S. W. Edwards, S. B. Stepaniants, S. Duenwald, R. Kleinhanz, A. R. Jones, D. D. Shoemaker, and R. B Stoughton. Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics*, pages 956–65., 2003.

[Heyer *et al.*, 1999]L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*, pages 1106–15., 1999.

[Pan, 2002]W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatic*, 2002.

[Reiner *et al.*, 2003]A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, pages 368–75., 2003.

[Sabatti *et al.*, 2002]C. Sabatti, L. Rohlin, M. K. Oh, and J. C. Liao. Co-expression pattern from dna microarray experiments as a tool for operon prediction. *Nucleic Acids Res*, pages 2886–93., 2002.

[Sekowska *et al.*, 2001]A. Sekowska, S. Robin, J. J. Daudin, A. Henaut, and A. Danchin. Extracting biological information from dna arrays: an unexpected link between arginine and methionine metabolism in bacillus subtilis. *Genome Biol*, pages 0019.1–0019.12, 2001.

[Troyanskaya *et al.*, 2002]O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, pages 1454–61., 2002.

3 Vision d'ensemble de l'organisation chromosomique bactérienne grâce à la méta-analyse

3.1 Découverte d'une structure d'expression du chromosome bactérien

Chez les eucaryotes, les différents types de chromatine (eu ou hétérochromatine) influencent grandement l'expression des gènes. Chez les procaryotes, aucune structure similaire n'a été observée. Cependant, plusieurs indices tendent à prouver que l'expression des gènes varie selon leur localisation sur le chromosome [139]. L'ordre de compaction de l'ADN bactérien est identique à l'ADN des eucaryotes : environ 7 000 fois. Cependant, la densité en gènes des chromosomes bactériens est nettement plus élevée. La compaction de l'ADN devrait donc être beaucoup plus dynamique que celle des eucaryotes [139].

Par ailleurs, chez les bactéries, la transcription et la traduction de l'ARNm sont simultanées. Or, ces processus s'effectuent à la surface du nucléoïde, la machinerie de transcription et de réplication ne pouvant pas pénétrer à l'intérieur. L'avancée des techniques d'hybridation massives ouvre la voie à de nouvelles études de la structure chromosomique. Si des gènes ont des niveaux d'expression corrélés, ils devraient se retrouver en même temps à la surface du nucléoïde. L'image obtenue par l'utilisation de plusieurs jeux de données d'expression indépendants devrait correspondre à une vision indépendante des buts expérimentaux et relativement complète de la co-expression des gènes et donc de la structure du chromosome.

Nous avons étudié les corrélations d'expression des gènes chez deux bactéries modèles, *Bacillus subtilis* et *Escherichia coli*, afin de déterminer si l'on trouve effectivement des régularités d'expression pouvant conduire à la description de l'organisation potentielle du nucléoïde.

Notre étude a montré que, pour deux bactéries modèle *Bacillus subtilis* et *Escherichia coli*, la co-expression des gènes varie en fonction de la distance des gènes sur le chromosome. Nous avons identifié des corrélations longue distance surprenantes : les changements de niveaux d'expression de n'importe quel gène sont corrélés (positivement ou négativement) aux changements d'expression d'autres gènes localisés à de grandes distances déterminées. Cette propriété est valable quel que soit le gène et quelle que soit sa localisation sur le chromosome.

Nous avons également identifié des corrélations à petites distances qui suggèrent que la localisation des gènes co-exprimés correspond à des tours d'ADN sur la surface du nucléoïde (14-16 gènes).

Les corrélations longues distances ne correspondent pas aux domaines identifiés précédemment dans le nucléoide. Nous avons interprété nos résultats à partir d'un modèle de la structure du nucléoide sous la forme d'un solénoïde à deux types de spirales. Les grandes spirales correspondraient à de l'ADN exprimé décondensé tandis que les petites spirales seraient de l'ADN non exprimé condensé.

BMC Genomics

BioMed Central

Research article

Open Access

Decoding the nucleoid organisation of *Bacillus subtilis* and *Escherichia coli* through gene expression data

Anne-Sophie Carpentier*[1], Bruno Torrésani[2], Alex Grossmann[1] and Alain Hénaut[1]

Address: [1]Laboratoire Génome et Informatique, CNRS UMR 8116, Tour Evry2, 523 Place des Terrasses, 91034 Evry Cedex, France and [2]CMI, Université de Provence, 39 rue Joliot-Curie, 13453 Marseille cedex 13, France

Email: Anne-Sophie Carpentier* - carpentier@genopole.cnrs.fr; Bruno Torrésani - Bruno.Torresani@cmi.univ-mrs.fr; Alex Grossmann - grossman@genopole.cnrs.fr; Alain Hénaut - henaut@genopole.cnrs.fr

* Corresponding author

Published: 06 June 2005

BMC Genomics 2005, **6**:84 doi:10.1186/1471-2164-6-84

Received: 04 April 2005
Accepted: 06 June 2005

This article is available from: http://www.biomedcentral.com/1471-2164/6/84

Abstract

Background: Although the organisation of the bacterial chromosome is an area of active research, little is known yet on that subject. The difficulty lies in the fact that the system is dynamic and difficult to observe directly. The advent of massive hybridisation techniques opens the way to further studies of the chromosomal structure because the genes that are co-expressed, as identified by microarray experiments, probably share some spatial relationship. The use of several independent sets of gene expression data should make it possible to obtain an exhaustive view of the genes co-expression and thus a more accurate image of the structure of the chromosome.

Results: For both *Bacillus subtilis* and *Escherichia coli* the co-expression of genes varies as a function of the distance between the genes along the chromosome. The long-range correlations are surprising: the changes in the level of expression of any gene are correlated (positively or negatively) to the changes in the expression level of other genes located at well-defined long-range distances. This property is true for all the genes, regardless of their localisation on the chromosome.

We also found short-range correlations, which suggest that the location of these co-expressed genes corresponds to DNA turns on the nucleoid surface (14–16 genes).

Conclusion: The long-range correlations do not correspond to the domains so far identified in the nucleoid. We explain our results by a model of the nucleoid solenoid structure based on two types of spirals (short and long). The long spirals are uncoiled expressed DNA while the short ones correspond to coiled unexpressed DNA.

Background

As Lovett and Segall [1] point out in their meeting report on the recently held "Keystone Symposium on Bacterial Chromosomes", we know a lot about the bacterial DNA replication, recombination, repair and other aspects of cell biology, but still rather little about the organisation of bacterial chromosome. The difficulty lies in the fact that the system varies and is difficult to observe directly. A number of different techniques are being employed to

answer the problem. The following is meant to give a brief overview and has no claim to be exhaustive:

• **Cytology**-based approaches include the use of DNA fluorescence microscopy, optical sectioning and FISH (fluorescence *in situ* hybridisation). These techniques were applied in order to localise within the cell a set of chromosomal segments [2] or to see the relationship between the shapes of the nucleoid and the underlying arrangements of DNA [3].

• Cunha et al [4] approach the question from a **cytometric** point of view, in order to study the compaction and the internal dynamics of the nucleoid.

• **An example of a classical genetic approach** is the work by Valens et al [5] who have used a site-specific recombination system in order to reveal spatial proximities of distant DNA sites.

• Various **genomic approaches** have been adopted. Some authors, like Audit and Ouzounis [6], have taken a sequence-based point of view, in which they face the issue of gene localisation and orientation using 89 complete microbial chromosomes from eubacteria and archeabacteria. This approach leaves aside any physiology-based consideration.

• Other authors have examined the **physiological constraints** operating placed upon the cell in order to infer chromosomal structure. The idea is that genes which use the same type of resource (e.g. a particular tRNA pool) or which are involved in a part of metabolism that needs a particular environment (e.g. genes involved in sulphur metabolism which is highly sensitive to free radicals) should be in close proximity in the cell, even if they are far away on the chromosome [7,8].

The approaches mentioned above can be spilt in two groups: (i) large-scale analyses, aiming at deciphering the global chromosome organisation; (ii) small-scale analyses, which take a particular point of view (some genes or markers are chosen). The introduction of **microarrays** has added yet another way to study the chromosomal structure, allowing simultaneously the analysis on small and large scales [9]. Microarrays allow the measure of relative expression levels of the whole genome and therefore the identification of those genes that are co-expressed. Usually the co-expressions observations are used to elucidate the structure of operons and other regulatory structures, see for example [10,11].

The present work aims at understanding the nucleoid structure with the help of microarray data. As transcriptionally active DNA is located near the nucleoid surface or on DNA loops extending from the nucleoid [12], the co-expressed genes which are identified with microarrays probably share some spatial relationship.

However, microarrays give significant information only for those genes the level of expression of which varies across experiments. Consequently, the experimental conditions should be diversified in order to obtain a list of gene correlations as exhaustive as possible and thus an accurate image of the chromosomal structure. To this end, we gathered a number of currently available microarray data from the literature. The data were then pooled together, and treated as just one large data set. This "pooling of information" has already been carried out successfully from human expression data for a study of gene function [13], and from yeast or bacterial data for regulation studies [11,14].

We applied this method to two distant bacteria: *Escherichia coli* and *Bacillus subtilis*. Audit and Ouzounis [6] had the same approach, expecting that if observations made on one organism also hold true for the other, it would be reasonable to assume that the inferred chromosomal organisation is indeed a general characteristic of bacteria with double stranded, circular DNA.

Results

The aim of this work is to delineate how the co-expression intensities (correlations) of pairs of genes vary as a function of the inter-gene distance along the chromosome. The co-expression intensity for each couple of genes was evaluated with a non-parametric correlation: the Kendall tau [15,16] (see methods and figure 1 part 2) which depends only on the sign of the observed variation and not on its magnitude. Is is thus a "weaker" describer of the data than the linear correlation coefficient (also called Pearson coefficient of correlation) or the Spearman rank correlation coefficient. The Kendall tau points specifically to monotonic correlations. A high Kendall tau between two genes indicates that their levels of expression vary in the same way: when the expression level of the first gene increases, the expression level of the other one increases also.

Then the variation of the Kendall tau coefficient as a function of the distance between genes was measured with a standard linear autocorrelation function [15,16] (see methods and figure 1 part 3). The linear autocorrelation enables to point to regularities in a gene Kendall tau vector and therfore to regularities of expression correlated with particular inter-gene distances.

1. Normalised data

	Experimental condition 1	Experimental condition 2	Experimental condition 3
gene 1	-0.39	1.45	1.21
gene 2	0.65	-1.26	0.77
gene 3	-1.43	-0.48	-0.99
gene 4	1.17	0.29	-0.99

2. Matrix of Kendall tau

	gene 1	gene 2	gene 3	gene 4
gene 1	1	-1/3	1	-1/3
gene 2	-1/3	1	-1/3	-1/3
gene 3	1	-1/3	1	-1/3
gene 4	-1/3	-1/3	-1/3	1

3. Matrix of linear autocorrelations

	Inter-gene distance 1	Inter-gene distance 2	Inter-gene distance 3	Inter-gene distance 4
gene 1	-1	1	-1	1
gene 2	-1/3	-1/3	-1/3	1
gene 3	-1	1	-1	1
gene 4	-1/3	-1/3	-1/3	1

4. Averaged linear autocorrelation

Inter-gene distance 1	Inter-gene distance 2	Inter-gene distance 3	Inter-gene distance 4
-2/3	1/3	-2/3	1

Figure 1

Illustration of the methodology used in this study. Example of the results obtained on a hypothetical bacterial circular chromosome model of 4 genes. The gene expression intensities are measured in three experimental conditions. Part 1 is normalised data (mean equal to 0 variance equal to 1) according to experimental conditions. Part 2 is the matrix of Kendall tau (see methods). Part 3 is the autocorrelation matrix with inter-gene distances. Part 4 is the averaged linear autocorrelation.

Bacillus subtilis *regularities of co-expression across the genome*

The analysis of the *B. subtilis* transcription data was performed on a set of 262 experimental conditions gathered from eleven independent experiments measuring expression data over the whole genome. A global view of the regularities of co-expression was obtained by summing up the autocorrelation vectors of all the genes (see figure 1 part 4 and results in figure 2 -blue curve).

The averaged linear autocorrelation of changes in gene expression varies as a function of the inter-gene distance. The green curve in figure 2 corresponds to the averaged autocorrelation evaluated after random permutation of the gene positions on the chromosome. Here the variations are small and independent of the inter-gene distances. Those points where the autocorrelation (blue curve) departs from the random signal (green curve) correspond to couples of genes, for which changes in expression levels are statistically correlated (when the blue curve is above the green one) or anti-correlated (when the blue curve is below the green one).

The autocorrelation function shows regular oscillations at large scale, with maxima at a distance of 200, 650, 850, 1300, 1500 and 2050 genes and minima at a distance of

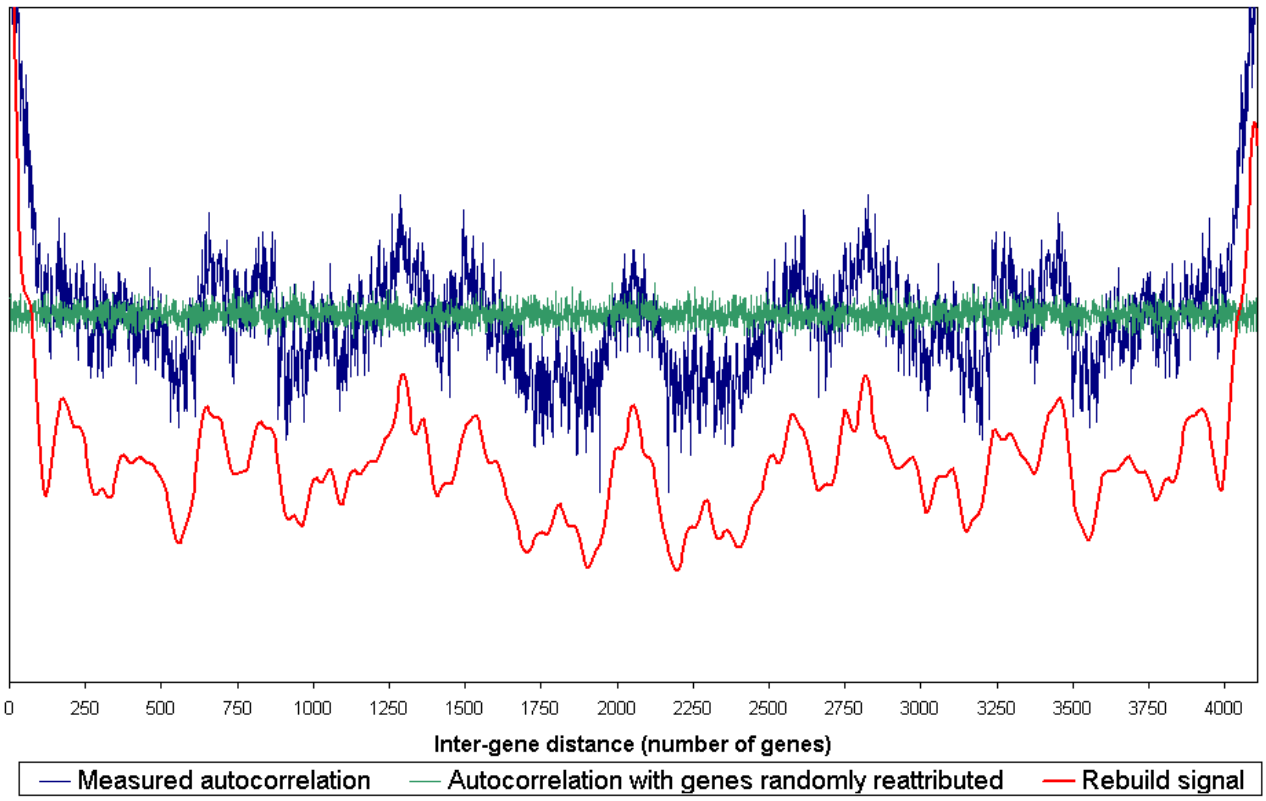


Figure 2
Long-range averaged autocorrelations in *B. subtilis*. To identify the regularities which are common to most of the genome, regardless of the genes localisation, the autocorrelation vectors of all the genes were summed (blue curve). This global signal shows the averaged autocorrelation regularities as a function of inter-gene distance. The green curve shows the averaged autocorrelation when the genes positions on the genome were randomly assigned. The red curve represents the resultant of four oscillations of periods 600 ± 55, 240 ± 21, 113 ± 21 and 60 ± 6 genes, which were estimated from the averaged autocorrelation deconvolution. The horizontal scale represents the distance between two genes (the difference of their ranks on the chromosome). The green, blue and red curves have the same vertical scale. The red curve was shifted for readability. Whereas the green signal shows no regularity, long-range correlations can be seen in the blue signal (maxima at *ca*. 200, 650, 850, 1300, 1500 and 2050 inter gene distance and minima at *ca*. 550, 900 and 1750–1950).

550, 900 and from 1750 to 1950 genes. Note that the inter-gene distance 2050 corresponds to diametrically opposite genes on the *B. subtilis* chromosome. The autocorrelation function can be seen as the resultant of four oscillations of periods 600 ± 55, 240 ± 21, 113 ± 21 and 60 ± 6 genes. This representation explains 85% of the autocorrelation oscillations (figure 2 – red curve).

The averaged autocorrelation was analysed on a smaller scale with an inter-gene distance comprised between 1 and 150 genes (figure 3 – blue curve). Closely spaced genes on the chromosome show changes in expression levels that are highly correlated. The averaged autocorrelation of two contiguous genes is 0.4. The low-scale autocorrelation can be decomposed into two regimes: (i) inter-gene distances between 1 and 5 (or 6) genes are characterised by a high and rapidly decaying autocorrelation; (ii) beyond a 6 inter-gene distance the autocorrelation shows a regular and slower decay with periodic oscillations of 14 to 15 genes (figure 3 – red curve). The autocorrelation merges with the noise background around an inter-gene distance of 100 genes (corresponding roughly to 100 kb).

The oscillations of the averaged autocorrelations of the 4108 *B. subtilis* genes shown in figure 2 may result (i) either from regularities specific to some genes or some regions; (ii) or from an overall property that would be shared by all the genes regardless of their positions on the

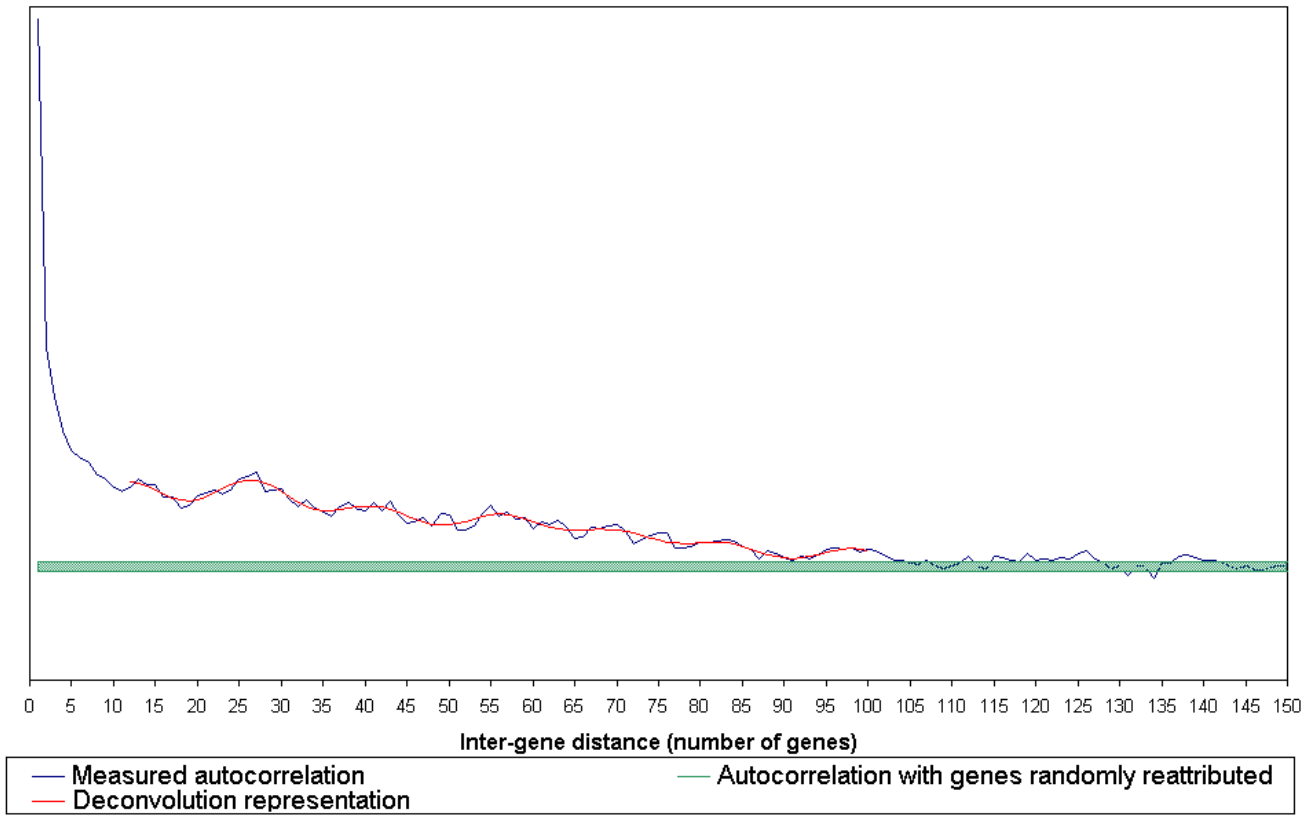


Figure 3

Short-range co-expression regularities in *B. subtilis*. To identify the regularities which are common to most of the genome, regardless of the genes localisation, the autocorrelation vectors of all the genes were summed (blue curve). This global signal shows the averaged autocorrelation regularities as a function of inter-gene distance. The green zone shows the averaged autocorrelation when the genes positions on the genome were randomly assigned (mean of the random signal ± the root mean square deviation). The horizontal scale represents the distance between two genes (the difference of their ranks on the chromosome). Neighbouring genes on the chromosome show highly correlated variations of expression levels. The averaged autocorrelation of two contiguous genes is 0.4. The signal can be decomposed into two parts: (i) inter-gene distances between 1 and 5–6 genes are characterised by a high autocorrelation, which drops steeply; (ii) beyond 6 genes the autocorrelation shows a regular and slower decrease. The autocorrelation merges with the background noise at an inter-gene distance of about 100 genes (similar to 100 kb). The autocorrelation decrease may be seen as the resultant of a linear decrease and 14.5 ± 1 genes period oscillations (red curve).

chromosome. In order to ascertain which hypothesis is the correct one, the sums of the autocorrelations of continuous groups of 10, 100 and 500 genes were calculated. All the curves obtained are highly similar (data shown for groups of 500 genes, figure 4). The peaks obtained with these groups of genes are identical to those found in the global signal. Hence they do not depend on any particular position on the genome: in other words, the results show that any gene A has its changes in expression level correlated with the changes in expression levels of those genes that are 200, 650, 850, 1300, 1500 and 2050 genes apart and anti-correlated with those that are 550, 900 and 1750–1950 genes apart. This property is independent of the position of gene A.

Escherichia coli *regularities of co-expression across the genome*

The same work was performed on *E. coli* with a data set of 106 experimental conditions. This data set is therefore smaller than that used for *B. subtilis*. In addition there are more missing data for *E. coli* than for *B. subtilis*.

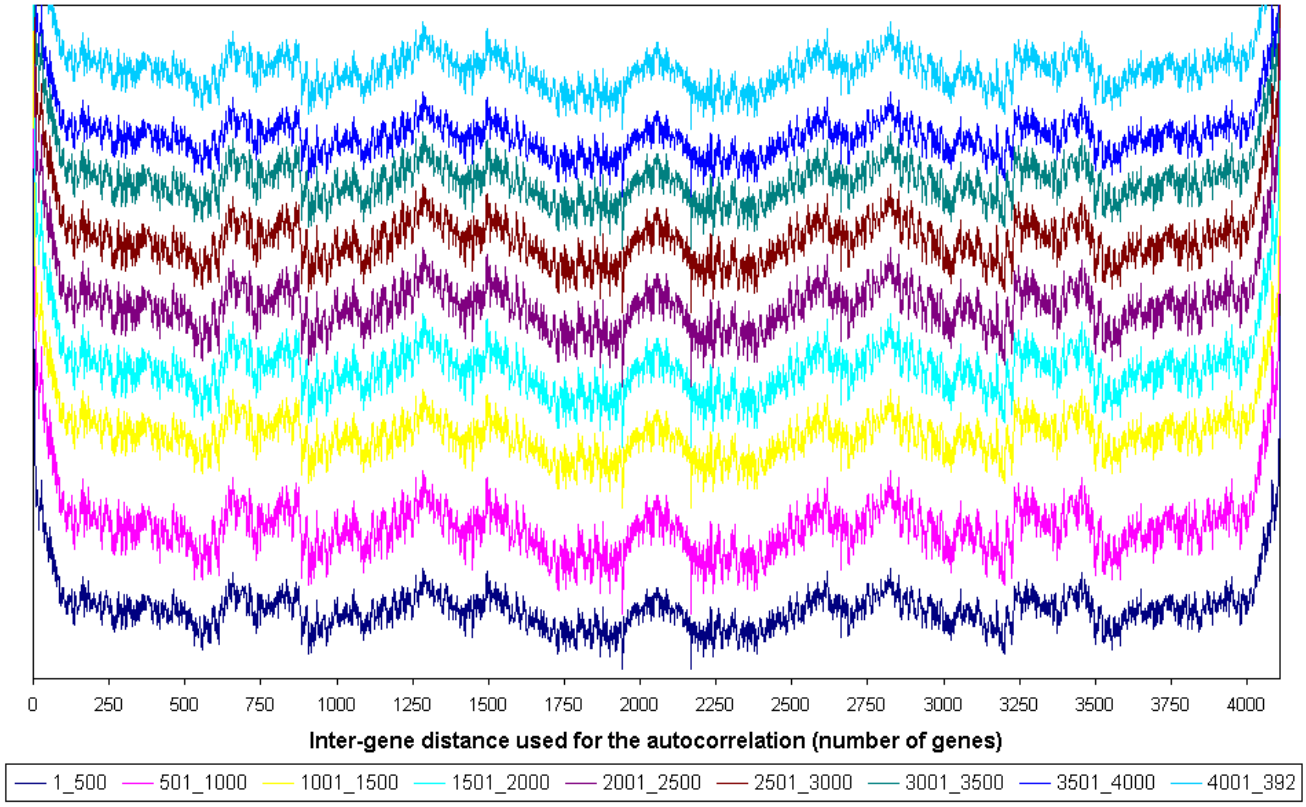


Figure 4

Partial sums of the autocorrelations in *B. subtilis*. To analyse if the discovered regularities depend on gene position, the autocorrelation vectors of groups of 500 genes were summed up (9 coloured curves). The horizontal scale represents the distance between two genes (the difference of their ranks on the chromosome). All the curves were vertically shifted for readability. The signals show the co-expression regularities according to inter-gene distance. Long-range periodicities are shared by all the signals regardless of the gene groups.

Figure 5 represents the variations of the averaged autocorrelation of all the genes as calculated with the actual gene positions (blue curve) and with random gene positions (green curve). The points where the autocorrelation (blue curve) departs from the random signal (green curve) correspond to couples of genes, the change in expression levels of which are correlated (when the blue curve is above the green one) or anti-correlated (when the blue curve is below the green one).

The main characteristics of figures 2 and 5 are similar. Both bacteria share the steep decay of the averaged autocorrelation curve for inter-gene distances lower than 100 genes and two maxima at a distance of 200 and 650 genes. However there are some differences between *B.subtilis* and *E.coli* for long-range peaks since some of them are shifted: maxima at 1300 and 1500 in *B.subtilis* correspond to peaks at 1100 and 1400 in *E.coli*, respectively. The minimum at 900 in *B.subtilis* is shifted to 850 in *E.coli*. Some peaks and troughs, however, are specific to one specie such as those located at 1380, 1700 and 2180 in *E.coli* and at 550, 850, 1750–1900 and 2050 in *B.subtilis*. Probably due to the greater number of missing data the autocorrelation function is noisier for *E.coli* than for *B.subtilis*.

Discussion

Comparison of our results to already published observations

What has already been observed

The present study of gene expression data from *B.subtilis* and *E.coli* has allowed us to confirm and extend some previously published observations:

• We show for both bacteria that closely spaced genes exhibit highly correlated expression levels. This correlation decreases rapidly with oscillations having a

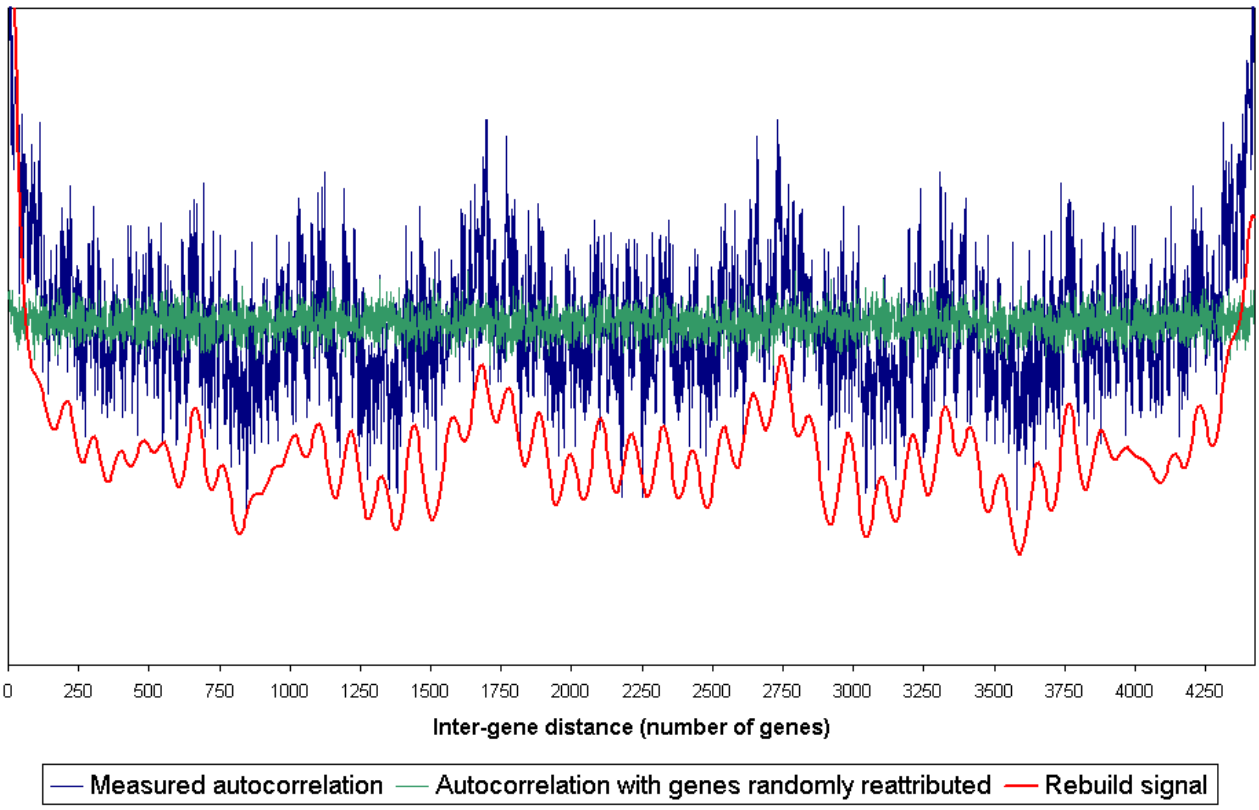


Figure 5
Long-range averaged autocorrelation in *E. coli*. To identify the regularities which are common to most of the genome, regardless of the genes localisation, the autocorrelation vectors of all the genes were summed (blue curve). This global signal shows the averaged autocorrelation regularities as a function of inter-gene distance. The green curve shows the averaged autocorrelation when the genes positions on the genome were randomly assigned. The red curve represents the resultant of two oscillations of periods 557 ± 30 and 100 ± 18 genes, which were estimated from the averaged autocorrelation deconvolution. The horizontal scale represents the distance between two genes (the difference of their ranks on the chromosome). The green and blue curves have the same vertical scale. The red one is on a scale, which is moved down for readability. Whereas the green signal shows no regularity, long-range periodicities can be seen in the blue signal (maxima at *ca.* 200, 650, 1100, 1400 and 1700 and minima at *ca.* 850, 1380 and 2180).

period of 14.5 ± 1 genes corresponding to 14.5 ± 1 kb. Short-range correlations are obvious in the study by Sabatti et al [11] of gene expression data from *E.coli*. Jeong et al [9] have also observed short-range correlations up to 16 kb in their analysis of expression changes during replication in various *E. coli* strains.

• In this work the averaged autocorrelation function for *E. coli* may be seen as the resultant of two main oscillations (with periods of 557 ± 30 kb and 100 ± 18 kb). In *B. subtilis* we observe four oscillations (with periods of 600 ± 55 kb, 240 ± 21 kb, 113 ± 21 kb and 60 ± 6 kb). Rocha et al [17] analysed the distribution of the genes involved in sulphur metabolism in the genome of *E.coli* and found a number of them to be clustered into statistically significant islands located 650 kb apart. In their study of transcriptional activities in *E.coli*, Jeong et al [9] have observed significant correlations for genes located 690 kb or 523 kb apart (depending on physiological conditions) together with a clump of periods around 115 kb.

New results

• We show here that the long-range and short-range correlations are similar in *E. coli* and *B. subtilis*. That the

observed regularities should be shared by two widely distant bacteria immediately suggests that it could be a property common to other bacteria as well.

• In addition, our results are indicative of an unexpected property that may well modify the current model of the nucleoid organisation: the changes in the level of expression of any gene are correlated (positively or negatively) to the changes in the expression level of other genes, located at well-defined long-range distances and regardless of their localisation on the chromosome in both organisms.

• The long-range periods of the autocorrelation function do not correspond to the 100 kb domain organisation, which may result from the control of topological constraints on the rotation of the double helix [12] and was observed in a study of the positions of genes that are controlled by a sequence-specific transcriptional regulator and the genes encoding this regulator [18]. They do not correspond either to the macro-domain of 1 or 2 Mb proposed by Niki et al [2] and by Valens et al [5]. As all the genes exhibit the same long-range correlations, the phenomenon cannot be explained by some process involving regulators. Conversely, the observations made by Jeong et al [9] may be the result of the general phenomenon observed in this study.

Our interpretation

Gene transcription can occur only on the nucleoid surface. Thus the expression correlations that we observed imply that the involved pairs of genes lies on this surface. However all the genes cannot be on the nucleoid surface at the same time. Therefore depending on the external conditions and/or physiological requirements of the cell, different groups of co-expressed genes should be accessible to the transcriptional machinery. Such constraint seems hardly compatible with an unstructured spatial organisation of the chromosome. Similarly a disordered or random packing is very unlikely to result in the significant periodicities described above. Rather, our observations suggest that the nucleoid must be packed in a fairly structured way.

Knowledge about the nucleoid and ribosomes sizes

The genome sizes of *E. coli* and *B. subtilis* are respectively 4.6 Mb (4425 genes encoding proteins) and 4.2 Mb (4108 genes encoding proteins). Half of the genes belong to an operon. The operons have an average size of three genes [12,19]. The nucleoid (the chromosome) shows up as a cylinder of approximate size of 0.5 × 0.7 μm [12,20]. Its circumference of 1.5 μm corresponds approximately to 16 kb of uncoiled DNA, or 16 genes. The diameter of a ribosome is 0.025 μm [21], hence 25 to 30 ribosomes can be juxtaposed along the cylinder length of 0.7 μm.

The possible chromosome configuration

We assume that the nucleoid structure consists of a solenoid with two types of spirals (figure 6):

• Large spirals of uncoiled DNA, containing the genes that are transcribed, that lie on the surface of the nucleoid and define its diameter.

• Small spirals of coiled untranscribed DNA that lie inside the nucleoid.

Cellular elements, in particular the ribosomes on the surface of the nucleoid, impose limits to the number of large expressed spirals. The distance between two large spirals cannot be shorter than the diameter of the ribosome; hence a maximum of 25 to 30 uncoiled DNA large spirals may stand on the nucleoid surface (see knowledge about the nucleoid and ribosomes sizes).

Short-range correlations show that contiguous co-expressed genes do not span more than 100 kb, hence no more than 6 large spirals. We can therefore assume that the average length of contiguous uncoiled DNA is 3 large spirals (see figure 6). This will make 8 to 10 groups of three consecutive large DNA spirals distributed along the chromosome.

Explanation of our results by this nucleoid representation

• The short-range correlations may be seen as resulting from two phenomena:

- The co-ordinated expression of the genes within operons. This explains the correlations in the expression of pairs of genes that are less than 5–6 genes apart.

- The presence of one or more consecutive DNA large spirals of approximately 16 genes on the nucleoid surface. The 14.5 ± 1genes period observed in the variations of the autocorrelation function points to those genes that belong to successive spirals and lie on a generatrix of the nucleoid cylinder.

• For long-range correlations we find 10 maxima in *E. coli* and 11 maxima in *B. subtilis*. These maxima probably result from groups of large DNA spirals on the nucleoid surface.

However, such a static representation of the nucleoid does explain neither the alternating pattern of maxima and minima nor their positions.

The dynamic of the nucleoid: a phenomenon, which is not fully explained

The dynamic of the nucleoid structure corresponds to the shift between small spirals of unexpressed coiled DNA to

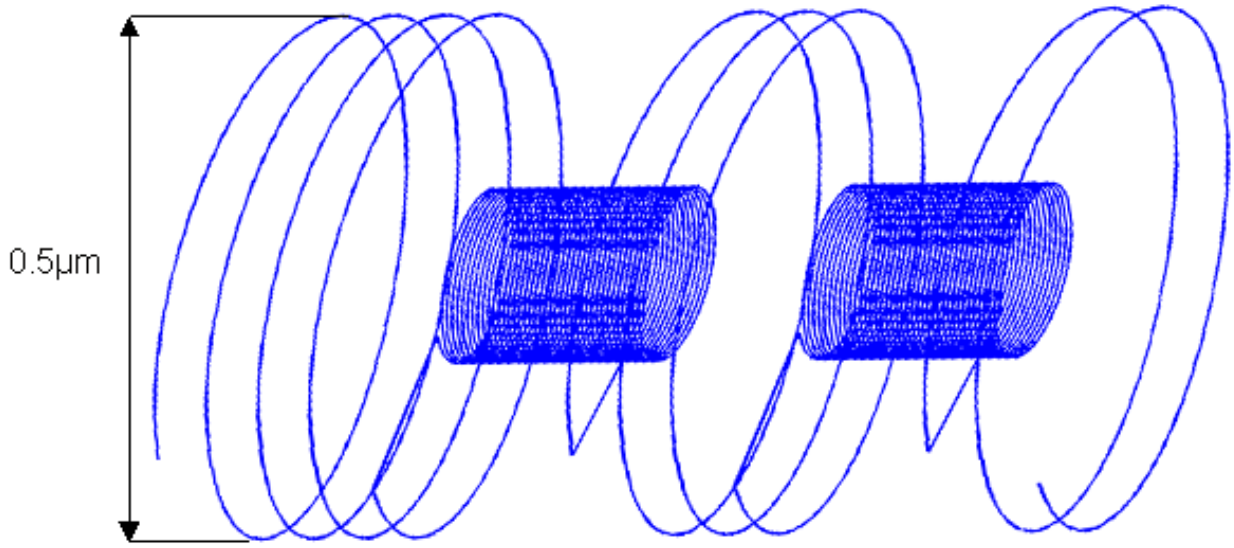


Figure 6
The possible chromosome configuration. We assume that the nucleoid structure consists of a solenoid with two types of spirals: • Large spirals of uncoiled DNA, containing the genes that are transcribed, that lie on the surface of the nucleoid and define its diameter (0.5 μm). • Small spirals of coiled untranscribed DNA that lie inside the nucleoid.

large spirals of expressed uncoiled DNA, and *vice-versa*. The large spirals are present only when there is effective transcription [22]. The transcription process can explain some of our observations:

- Long-range anticorrelations can result from coil-coiled DNA in small spirals next to large expressed spirals. It has been shown indeed that the opening of the double-stranded DNA during transcription leads to waves of compression of those regions of the chromosome that are close to the transcribed DNA [23]. It can therefore be speculated that the expression of the genes in large spirals leads to the super-coiling of the neighbouring small spirals, hence to the impossibility of opening its DNA and to its transcription.

The pattern of maxima is more difficult to explain since it does not correspond to multiples of a single inter-gene distance. In the case of *B. subtilis* for example, the maxima are at inter-gene distances that are multiples of 650 and multiples of 650 plus 200 (200; 650, 850; 1300, 1500). We speculate that this pattern is a consequence of the dynamic of the nucleoid structure but we currently have no explanation for it. Current work is in progress to try to explain the maxima and minima of the correlation function, which is reminiscent of a beat phenomenon between two stable waves that could be generated by the transcription process.

Conclusion

The analysis of gene expression data compendium provided information on the nucleoid organisation in circular double stranded DNA bacteria. Our results confirm and complete other observations like those obtained by microscopy. Co-expression variations of neighbouring genes on the chromosome suggest that large DNA spires of 14 to 16 genes length stay on the nucleoid surface. This estimation of a large spire length corresponds to the estimation by microscopy of the nucleoid circumference. The contiguous DNA on the nucleoid surface does not exceed around 100 genes (which is equivalent to 100 kb). This segment is organised in several large spirals of 14 to 16 genes.

The long-range correlation pattern is more surprising: the changes in level of expression for any gene are correlated (positively or negatively) to the changes in expression level of genes, located at well-defined long-range distances independently of their location on the chromosome. This original observation is based on the analysis of

several independent sets of gene expression data, which put together a great variety of physiological conditions. However the long-range correlations do not correspond to the domains identified so far in the nucleoid. We are currently exploring a model where the long-range correlations could result from a beat phenomenon between compression and decompression waves generated by the transcription process.

Methods

Data used and normalisation

The microarray data sets have been downloaded from the following websites.

Bacillus subtilis

- Helmann et al [24] at the Stanford microarray database http://genome-www5.stanford.edu/MicroArray/SMD/

- Yoshida et al[25,26], Ogura et al [27,28], Kobayashi et al [29], Asai et al [30], Doan et al [31], Molle et al [32], and Watanabe et al [33] at KEGG expression database http://www.genome.jp/dbget-bin/get_htext?Exp_DB+-e+L+C+-s+F+-f+F+C

- Jarmer et al [34] at the Center for Biological Sequence analysis site http://www.cbs.dtu.dk/~steen/Bacillus.html

Escherichia coli

- Mori et al [35] at KEGG expression database http://www.genome.ad.jp/dbget-bin/get_htext?Exp_DB+-e+L+C+-s+F+-f+F+C

- Newton et al [36] at http://www.stat.wisc.edu/~newton/papers/abstracts/btr139a.html

The data were normalised (mean equal to 0 and variance equal to 1) according to the experimental conditions (figure 1 part 1). They were concatenated for each organism leading to a file of gene expression levels made of 262 experimental conditions for *B. subtilis* and 106 experimental conditions for *E. coli*.

Estimation of the correlations and the regularities (figure 1)

The aim of this article is to observe how gene co-expressions vary as a function of the inter-gene distance.

1. For each organism the co-expression among each pair of genes is evaluated with a non-parametric correlation: the Kendall tau [15,16] (figure 1 part 2).

To define the Kendall tau τ, we start with the N data points (xi, yi), the expression levels of the genes x and y in the experimental condition i, respectively. Considering all the $1/2N(N - 1)$ pairs of data points (xi, yi) (xj, yj), we call a pair "concordant" if the differences $(xi-xj)$ and $(yi-yj)$ have the same sign and "discordant" if the differences have opposite signs. If $(xi-xj)$ is equal to zero, we call the pair an "extra y pair." If $(yi-yj)$ is equal to zero, we call the pair an "extra x pair." If both $(xi-xj)$ and $(yi-yj)$ are equal to zero the pair is ignored. Kendall's tau τ is the following simple combination of these various counts:

$$\tau = \frac{concordant - discordant}{\sqrt{concordant + discordant + extra\_y} \times \sqrt{concordant + discordant + extra\_x}}$$

2. For each gene, we evaluate its distances from those other genes, the expression levels of which vary simultaneously. The variations of co-expression according to inter-gene distance (figure 1 part 3) are evaluated with the linear autocorrelation [16] on the gene's Kendall tau vector.

The autocorrelation for an inter-gene distance of j is calculated as followed:

$$corr(j) = \frac{1}{N}\left(\sum_{x=1}^{N} (\gamma_x - \overline{\gamma})(\gamma_{x+j} - \overline{\gamma}) \right) \Big/ \sum_{x=1}^{N} (\gamma_x - \overline{\gamma})^2$$

with γ the Kendall tau vector of a gene and $\overline{\gamma}$ the mean of γ, N the number of genes

Note that the bacterial chromosome is circular, so there is no boundary problem. Note that the distance between two genes used in this article is the difference their ranks on the chromosome (approximately equivalent to the number of kb).

Signal deconvolution and estimation of the periodicities

The variation of co-expression according to the inter-gene distance is a superimposition of several periodicities (from small to large scale). To identify these periodicities the averaged autocorrelation signal was deconvoluated with Peakfit 4.06 (Jandel Scientific, San Rafael, CA). The percentage of the autocorrelation that this representation explains is calculated as follow:

$$z = 1 - \sum_{j=1}^{N} (x_j - \gamma_j)\dagger / \sum_{j=1}^{N} (\gamma_j - \overline{\gamma})\dagger$$

with γ the autocorrelation vector and x the signal generated by the sum of the deconvolution periodicities and N the number of genes.

Authors' contributions

ASC collected the data, performed the statistical analyses and drafted the manuscript. AG and BT participated in the statistical analysis. AH conceived the study, participated in its analysis and coordination. All authors participated to the elaboration of the model, read and approved the final manuscript.

Acknowledgements

The authors wish to thank A. Riva and J.-L. Risler for critical reading of the manuscript. The authors are indebted to Infobiogen for the disk space and calculation time provided on their servers. A.-S. Carpentier was supported by a FCPR of the Ministère de l'Agriculture.

References

1. Lovett ST, Segall AM: **New views of the bacterial chromosome.** *EMBO Rep* 2004, **5:**860-864.
2. Niki H, Yamaichi Y, Hiraga S: **Dynamic organization of chromosomal DNA in Escherichia coli.** *Genes Dev* 2000, **14:**212-223.
3. Zimmerman SB: **Underlying regularity in the shapes of nucleoids of Escherichia coli: implications for nucleoid organization and partition.** *J Struct Biol* 2003, **142:**256-265.
4. Cunha S, Woldringh CL, Odijk T: **Polymer-mediated compaction and internal dynamics of isolated Escherichia coli nucleoids.** *J Struct Biol* 2001, **136:**53-66.
5. Valens M, Penaud S, Rossignol M, Cornet F, Boccard F: **Macrodomain organization of the Escherichia coli chromosome.** *Embo J* 2004, **23:**4330-4341.
6. Audit B, Ouzounis CA: **From genes to genomes: universal scale-invariant properties of microbial chromosome organisation.** *J Mol Biol* 2003, **332:**617-633.
7. Nitschke P, Guerdoux-Jamet P, Chiapello H, Faroux G, Henaut C, Henaut A, Danchin A: **Indigo: a World-Wide-Web review of genomes and gene functions.** *FEMS Microbiol Rev* 1998, **22:**207-227.
8. Danchin A, Guerdoux-Jamet P, Moszer I, Nitschke P: **Mapping the bacterial cell architecture into the chromosome.** *Philos Trans R Soc Lond B Biol Sci* 2000, **355:**179-190.
9. Jeong KS, Ahn J, Khodursky AB: **Spatial patterns of transcriptional activity in the chromosome of Escherichia coli.** *Genome Biol* 2004, **5:**R86.
10. Steinhauser D, Junker BH, Luedemann A, Selbig J, Kopka J: **Hypothesis-driven approach to predict transcriptional units from gene expression data.** *Bioinformatics* 2004, **20:**1928-1939.
11. Sabatti C, Rohlin L, Oh MK, Liao JC: **Co-expression pattern from DNA microarray experiments as a tool for operon prediction.** *Nucleic Acids Res* 2002, **30:**2886-2893.
12. Pettijohn DE: **The Nucleoid.** In *Escherichia coli and Salmonella Cellular and Molecular Biology* Second Edition edition. Edited by: F.C. Neidhardt . Washington, DC, ASM Press; 1999.
13. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14:**1085-1094.
14. Allocco DJ, Kohane IS, Butte AJ: **Quantifying the relationship between co-expression, co-regulation and gene function.** *BMC Bioinformatics* 2004, **5:**18.
15. Press WH: **Kendall's Tau.** In *Numerical recipes in C : the art of scientific computing* 2nd edition. Cambridge, Cambridge University Press,; 1992:643-645.
16. Kendall MG, Ord JK, Stuart A: **Stationay time-series.** In *The advanced theory of statistics 3 Design and analysis, and time series Volume 3*. 4th edition. London, Griffin; 1983.
17. Rocha EP, Sekowska A, Danchin A: **Sulphur islands in the Escherichia coli genome: markers of the cell's architecture?** *FEBS Lett* 2000, **476:**8-11.
18. Kepes F: **Periodic transcriptional organization of the E.coli genome.** *J Mol Biol* 2004, **340:**957-964.
19. Willenbrock H, Ussery DW: **Chromatin architecture and gene expression in Escherichia coli.** *Genome Biol* 2004, **5:**252.
20. Donachie W, Begg KJ: **Cell Length, Nucleoid Separation, and Cell Division of Rod-Shaped and Spherical Cells of Escherichia coli.** *J Bacteriology* 1989, **171:**4633-4639.
21. Noller HF, Nomua M: **Ribosomes.** In *Escherichia coli and Salmonella Cellular and Molecular Biology* Second Edition edition. Edited by: F.C. Neidhardt . Washington, DC, ASM Press; 1999.
22. Murphy LD, Zimmerman SB: **Hypothesis: the RNase-sensitive restraint to unfolding of spermidine nucleoids from Escherichia coli is composed of cotranslational insertion linkages.** *Biophys Chem* 2002, **101-102:**321-331.
23. Krasilnikov AS, Podtelezhnikov A, Vologodskii A, Mirkin SM: **Large-scale effects of transcriptional DNA supercoiling in vivo.** *J Mol Biol* 1999, **292:**1149-1160.
24. Helmann JD, Wu MF, Gaballa A, Kobel PA, Morshedi MM, Fawcett P, Paddon C: **The global transcriptional response of Bacillus subtilis to peroxide stress is coordinated by three transcription factors.** *J Bacteriol* 2003, **185:**243-253.
25. Yoshida K, Kobayashi K, Miwa Y, Kang CM, Matsunaga M, Yamaguchi H, Tojo S, Yamamoto M, Nishi R, Ogasawara N, Nakayama T, Fujita Y: **Combined transcriptome and proteome analysis as a powerful approach to study genes under glucose repression in Bacillus subtilis.** *Nucleic Acids Res* 2001, **29:**683-692.
26. Yoshida K, Yamaguchi H, Kinehara M, Ohki YH, Nakaura Y, Fujita Y: **Identification of additional TnrA-regulated genes of Bacillus subtilis associated with a TnrA box.** *Mol Microbiol* 2003, **49:**157-165.
27. Ogura M, Yamaguchi H, Yoshida K, Fujita Y, Tanaka T: **DNA microarray analysis of Bacillus subtilis DegU, ComA and PhoP regulons: an approach to comprehensive analysis of B.subtilis two- component regulatory systems.** *Nucleic Acids Res* 2001, **29:**3804-3813.
28. Ogura M, Yamaguchi H, Kobayashi K, Ogasawara N, Fujita Y, Tanaka T: **Whole-genome analysis of genes regulated by the Bacillus subtilis competence transcription factor ComK.** *J Bacteriol* 2002, **184:**2344-2351.
29. Kobayashi K, Ogura M, Yamaguchi H, Yoshida K, Ogasawara N, Tanaka T, Fujita Y: **Comprehensive DNA microarray analysis of Bacillus subtilis two- component regulatory systems.** *J Bacteriol* 2001, **183:**7365-7370.
30. Asai K, Yamaguchi H, Kang CM, Yoshida K, Fujita Y, Sadaie Y: **DNA microarray analysis of Bacillus subtilis sigma factors of extracytoplasmic function family.** *FEMS Microbiol Lett* 2003, **220:**155-160.
31. Doan T, Servant P, Tojo S, Yamaguchi H, Lerondel G, Yoshida K, Fujita Y, Aymerich S: **The Bacillus subtilis ywkA gene encodes a malic enzyme and its transcription is activated by the YufL/YufM two-component system in response to malate.** *Microbiology* 2003, **149:**2331-2343.
32. Molle V, Nakaura Y, Shivers RP, Yamaguchi H, Losick R, Fujita Y, Sonenshein AL: **Additional targets of the Bacillus subtilis global regulator CodY identified by chromatin immunoprecipitation and genome-wide transcript analysis.** *J Bacteriol* 2003, **185:**1911-1922.
33. Watanabe S, Hamano M, Kakeshita H, Bunai K, Tojo S, Yamaguchi H, Fujita Y, Wong SL, Yamane K: **Mannitol-1-phosphate dehydrogenase (MtlD) is required for mannitol and glucitol assimilation in Bacillus subtilis: possible cooperation of mtl and gut operons.** *J Bacteriol* 2003, **185:**4816-4824.
34. Jarmer H, Berka R, Knudsen S, Saxild HH: **Transcriptome analysis documents induced competence of Bacillus subtilis during nitrogen limiting conditions.** *FEMS Microbiol Lett* 2002, **206:**197-200.
35. Mori H, Horiuchi T, Isono K, Wada C, Kanaya S, Kitagawa M, Ara T, Ohshima H: **[Post sequence genome analysis of Escherichia coli].** *Tanpakushitsu Kakusan Koso* 2001, **46:**1977-1985.
36. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comput Biol* 2001, **8:**37-52.

3.2 Validité de cette organisation selon la taille du chromosome bactérien

Les observations de l'article précédent mènent à l'idée que le repliement du chromosome bactérien pourrait être un phénomène physique sans doute lié à l'élasticité de l'ADN ainsi qu'à la taille du chromosome. Afin d'affiner ces observations, il est donc nécessaire d'avoir des observations pour des tailles de génomes différents. Or, *E. coli* et *B. subtilis* présentent un génome de taille comparable aux alentours de 4 000 pb.

Nous avons donc recherché une bactérie avec un ou des chromosomes circulaires de taille différant de 4 000 pb pour laquelle on dispose de données de transcriptome. *S. meliloti* présente l'intérêt de présenter un chromosome circulaire de l'ordre de 3 300 pb et deux mégaplasmides de 1 600 et 1 300 pb.

Nous avons donc reproduit la même étude sur ces trois éléments génomiques circulaires. Les résultats obtenus chez *E. coli* et *B. subtilis* ont été confirmés sur le chromosome bactérien de *S. meliloti* : on observe des régularités à grande échelle ainsi que la décroissance de l'autocorrélation jusqu'à une distance inter-génique de 100 gènes.

Les signaux chez les plasmides semblent également montrer des régularités à grande échelle de l'expression des gènes. Afin d'approfondir notre modèle, il s'avère nécessaire d'étudier plus en profondeur les régularités de corrélation et d'anti-corrélation dans les différents chromosomes et plasmides. Une question en suspens est de savoir si le comportement des plasmides est identique à celui d'un chromosome sur le plan des régularités de corrélation. Pour approfondir cette question, nous sommes à la recherche de données d'expression d'autres bactéries et notamment d'autres plasmides.

4 Discussion/Conclusion

Même si de nombreuses méthodes d'analyse sont utilisées dans le domaine du transcriptome, aucune n'est réellement spécifique à ce domaine. Beaucoup de méthodes statistiques utilisées sur les données de puces à ADN ne sont que des transpositions de méthodes développées dans d'autres domaines, certaines en sont des adaptations. Si peu de nouveautés peuvent être identifiées dans les méthodologies, les données de puces à ADN restent une aubaine pour l'application de méthodes statistiques. Le transcriptome ouvre un nouveau champ vaste d'applications des statistiques avec de nombreuses données désormais en libre disposition. Parallèlement, les statistiques sont un passage obligé dans l'analyse du transcriptome. Du fait du nombre important de mesures simultanées obtenues par les puces, seules des analyses statistiques permettent d'aborder des conclusions robustes. Néanmoins, l'analyse du transcriptome ne peut se faire qu'en combinant une approche statistique avec des données biologiques : d'où le nom de biostatistiques. C'est ce que nous avons tenté de faire avec la mise au point d'un protocole de comparaison des résultats de méthodes statistiques fondé sur une information biologique, les opérons,.

Si le transcriptome n'est pas le domaine privilégié pour le développement de méthodes statistiques, il est une réelle opportunité pour la biologie. Tout comme l'acquisition de la séquence de génome complet d'organisme a ouvert, il y a une dizaine d'année, la possibilité d'études étendues sur des mécanismes génétiques, le transcriptome ouvre, désormais, de nouveaux domaines d'études de la biologie. Il permet également d'approfondir des études difficiles à mener avec les techniques existantes auparavant. C'est le cas pour l'étude de la structure du chromosome bactérien. De nombreuses études ont tenté d'élucider le problème de l'organisation chromosomique des bactéries. Elles présentent toutes la caractéristique d'être des études indirectes. En effet, l'organisation dynamique du chromosome se prête peu à l'observation directe [140]. Différents types d'études ont été appliquées :

- cytologique. Elle comprennent la recherche de la localisation de différents segments chromosomiques dans la bactérie [141] ou l'étude de la forme du nucléoïde en fonction de la vitesse de croissance des bactéries [142]. Ces études permettent d'observer grossièrement la forme du nucléoïde grâce à différents marqueurs répartis sur le chromosome. Cependant, ces techniques ne permettent pas une étude fine exhaustive du fait du nombre limité de marqueurs.

- génétique. Une étude génétique classique [143] vise à identifier les proximités spatiales de segments chromosomiques distants grâce à un système de recombinaison. Cette étude n'est pas exhaustive puisqu'elle dépend du nombre limité de constructions génétiques effectuées. Par ailleurs, l'organisation « naturelle » du chromosome s'en retrouve modifiée.
- génomique. Les études génomiques reposent soit sur la composition en nucléoïdes, soit sur la disposition et l'orientation des gènes [144]. Le but est de rechercher des régularités de structure dans les données. Cette approche correspond à une étude statique des données indépendante de la physiologie de la bactérie.

Toutes ces approches présentent des limites, certaines n'étudient que la structure à petite échelle, de façon incomplète (fonction des marqueurs utilisé) ou statique.

Le transcriptome permet d'aborder la structure chromosomique bactérienne d'un autre point de vue. Il donne une vision dynamique de l'expression de l'ensemble des gènes de la bactérie et permet, ainsi, d'avoir un ensemble complet de marqueurs sur le chromosome. Il ouvre donc la possibilité d'avoir une vision complète et relativement dynamique de l'organisation du chromosome bactérien, ce qui était jusqu'alors impossible.

Le transcriptome, comme la génomique, ouvre de nouveaux horizons dans la biologie. De nouveaux domaines, non abordés dans ce manuscrit, verront sans doute le jour grâce aux données accumulées sur l'expression des gènes. Il est cependant difficile d'en déterminer leur étendue. Il est néanmoins raisonnable de reconnaître que ces données d'expression de l'ensemble des gènes sont une nouvelle étape dans la compréhension du fonctionnement du système dynamique qu'est la cellule. La vision ouverte par le transcriptome est beaucoup plus dynamique et fonctionnelle que l'est la vision issue de la génomique. Au fur et à mesure des progrès techniques, on se rapproche ainsi des réseaux fonctionnels de l'organisme. Certaines personnes pensent que la prochaine étape sera la connaissance de l'ensemble des protéines avec leurs différentes formes présentes au sein d'une cellule à un temps donné. Ces protéines représentent l'ensemble de la machinerie cellulaire, base des différents métabolismes présents. Cependant, le protéome ne représente pas l'ultime étape de la connaissance en biologie. Il serait plus juste de dire que le transcriptome et le protéome sont

complémentaires. En effet, les protéines sont stockées au sein de la cellule, ce que ne sont pas les ARNm. Les protéines donnent donc l'état cellulaire à un moment donné tandis que le transcriptome renseigne sur la dynamique du système et des réponses biologiques aux différents facteurs extérieurs.

5 Bibliographie

1. Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 33-7.
2. Lashkari, D.A., et al., *Yeast microarrays for genome wide parallel genetic and gene expression analysis.* Proc Natl Acad Sci U S A, 1997. **94**(24): p. 13057-62.
3. Vrana, K.E., W.M. Freeman, and M. Aschner, *Use of microarray technologies in toxicology research.* Neurotoxicology, 2003. **24**(3): p. 321-32.
4. Vingron, M., *Bioinformatics needs to adopt statistical thinking.* Bioinformatics, 2001. **17**(5): p. 389-90.
5. Jeong, K.S., J. Ahn, and A.B. Khodursky, *Spatial patterns of transcriptional activity in the chromosome of Escherichia coli.* Genome Biol, 2004. **5**(11): p. R86.
6. Templin, M.F., et al., *Protein microarray technology.* Trends Biotechnol, 2002. **20**(4): p. 160-6.
7. Alwine, J.C., D.J. Kemp, and G.R. Stark, *Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5350-4.
8. Liang, P. and A.B. Pardee, *Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction.* Science, 1992. **257**(5072): p. 967-71.
9. Velculescu, V.E., et al., *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.
10. Pollock, J.D., *Gene expression profiling: methodological challenges, results, and prospects for addiction research.* Chem Phys Lipids, 2002. **121**(1-2): p. 241-56.
11. van Hal, N.L., et al., *The application of DNA microarrays in gene expression analysis.* J Biotechnol, 2000. **78**(3): p. 271-80.
12. Kafatos, F.C., C.W. Jones, and A. Efstratiadis, *Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure.* Nucleic Acids Res, 1979. **7**(6): p. 1541-52.
13. Moody, D.E., *Genomics techniques: An overview of methods for the study of gene expression.* J. Anim. Sci., 2001. **79**: p. E128-E135.
14. Ekins, R., F. Chu, and J. Micallef, *High specific activity chemiluminescent and fluorescent markers: their potential application to high sensitivity and 'multi-analyte' immunoassays.* J Biolumin Chemilumin, 1989. **4**(1): p. 59-78.
15. Ekins, R.P., *Multi-analyte immunoassay.* J Pharm Biomed Anal, 1989. **7**(2): p. 155-68.
16. Ekins, R.P., F. Chu, and E. Biggart, *Multispot, multianalyte, immunoassay.* Ann Biol Clin (Paris), 1990. **48**(9): p. 655-66.
17. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
18. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
19. Mockler, T.C., et al., *Applications of DNA tiling arrays for whole-genome analysis.* Genomics, 2005. **85**(1): p. 1-15.
20. Fodor, S.P., et al., *Light-directed, spatially addressable parallel chemical synthesis.* Science, 1991. **251**(4995): p. 767-73.
21. Southern, E.M., U. Maskos, and J.K. Elder, *Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models.* Genomics, 1992. **13**(4): p. 1008-17.
22. Hacia, J.G., *Resequencing and mutational analysis using oligonucleotide microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 42-7.
23. Pease, A.C., et al., *Light-generated oligonucleotide arrays for rapid DNA sequence analysis.* Proc Natl Acad Sci U S A, 1994. **91**(11): p. 5022-6.
24. Davignon, L., et al., *Use of resequencing oligonucleotide microarrays for identification of Streptococcus pyogenes and associated antibiotic resistance determinants.* J Clin Microbiol, 2005. **43**(11): p. 5690-5.
25. Lipshutz, R.J., et al., *High density synthetic oligonucleotide arrays.* Nat Genet, 1999. **21**(1 Suppl): p. 20-4.
26. Ahrendt, S.A., et al., *Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array.* Proc Natl Acad Sci U S A, 1999. **96**(13): p. 7382-7.
27. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077-82.

28. Winzeler, E.A., et al., *Direct allelic variation scanning of the yeast genome.* Science, 1998. **281**(5380): p. 1194-7.
29. Shi, M.M., *Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies.* Clin Chem, 2001. **47**(2): p. 164-72.
30. Renn, S.C., N. Aubin-Horth, and H.A. Hofmann, *Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray.* BMC Genomics, 2004. **5**(1): p. 42.
31. Kallioniemi, A., et al., *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.* Science, 1992. **258**(5083): p. 818-21.
32. Pollack, J.R., et al., *Genome-wide analysis of DNA copy-number changes using cDNA microarrays.* Nat Genet, 1999. **23**(1): p. 41-6.
33. Rodriguez, B.A. and T.H. Huang, *Tilling the chromatin landscape: emerging methods for the discovery and profiling of protein-DNA interactions.* Biochem Cell Biol, 2005. **83**(4): p. 525-34.
34. Hanlon, S.E. and J.D. Lieb, *Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays.* Curr Opin Genet Dev, 2004. **14**(6): p. 697-705.
35. Iyer, V.R., et al., *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.* Nature, 2001. **409**(6819): p. 533-8.
36. Cawley, S., et al., *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.* Cell, 2004. **116**(4): p. 499-509.
37. Kapranov, P., et al., *Large-scale transcriptional activity in chromosomes 21 and 22.* Science, 2002. **296**(5569): p. 916-9.
38. Kampa, D., et al., *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.* Genome Res, 2004. **14**(3): p. 331-42.
39. Yamada, K., et al., *Empirical analysis of transcriptional activity in the Arabidopsis genome.* Science, 2003. **302**(5646): p. 842-6.
40. Johnson, J.M., et al., *Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments.* Trends Genet, 2005. **21**(2): p. 93-102.
41. Castle, J., et al., *Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing.* Genome Biol, 2003. **4**(10): p. R66.
42. Lee, C. and M. Roy, *Analysis of alternative splicing with microarrays: successes and challenges.* Genome Biol, 2004. **5**(7): p. 231.
43. Mata, J., S. Marguerat, and J. Bahler, *Post-transcriptional control of gene expression: a genome-wide perspective.* Trends Biochem Sci, 2005. **30**(9): p. 506-14.
44. Wheeler, D.B., A.E. Carpenter, and D.M. Sabatini, *Cell microarrays and RNA interference chip away at gene function.* Nat Genet, 2005. **37 Suppl**: p. S25-30.
45. Kutalik, Z., et al., *Advanced significance analysis of microarray data based on weighted resampling: a comparative study and application to gene deletions in Mycobacterium bovis.* Bioinformatics, 2004. **20**(3): p. 357-63.
46. Le Naour, F., et al., *Profiling changes in gene expression during differentiation and maturation of monocyte-derived dendritic cells using both oligonucleotide microarrays and proteomics.* J Biol Chem, 2001. **276**(21): p. 17920-31.
47. Clark, E.A., et al., *Genomic analysis of metastasis reveals an essential role for RhoC.* Nature, 2000. **406**(6795): p. 532-5.
48. DeRisi, J., et al., *Use of a cDNA microarray to analyse gene expression patterns in human cancer.* Nat Genet, 1996. **14**(4): p. 457-60.
49. Hedenfalk, I., et al., *Gene-expression profiles in hereditary breast cancer.* N Engl J Med, 2001. **344**(8): p. 539-48.
50. Yang, G. and S. Komatsu, *Microarray and proteomic analysis of brassinosteroid- and gibberellin-regulated gene and protein expression in rice.* Genomics Proteomics Bioinformatics, 2004. **2**(2): p. 77-83.
51. Eulgem, T., *Regulation of the Arabidopsis defense transcriptome.* Trends Plant Sci, 2005. **10**(2): p. 71-8.
52. Sekowska, A., et al., *Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in Bacillus subtilis.* Genome Biol, 2001. **2**(6): p. 0019.1-0019.12.
53. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

54. Niehrs, C. and N. Pollet, *Synexpression groups in eukaryotes.* Nature, 1999. **402**(6761): p. 483-7.
55. Chu, S., et al., *The transcriptional program of sporulation in budding yeast.* Science, 1998. **282**(5389): p. 699-705.
56. Cho, R.J., et al., *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell, 1998. **2**(1): p. 65-73.
57. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol Biol Cell, 1998. **9**(12): p. 3273-97.
58. Schaffer, R., et al., *Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis.* Plant Cell, 2001. **13**(1): p. 113-23.
59. Cho, R.J., et al., *Transcriptional regulation and function during the human cell cycle.* Nat Genet, 2001. **27**(1): p. 48-54.
60. Whitfield, M.L., et al., *Identification of genes periodically expressed in the human cell cycle and their expression in tumors.* Mol Biol Cell, 2002. **13**(6): p. 1977-2000.
61. Gasch, A.P., et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 2000. **11**(12): p. 4241-57.
62. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.
63. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
64. Garber, M.E., et al., *Diversity of gene expression in adenocarcinoma of the lung.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13784-9.
65. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat Med, 2002. **8**(8): p. 816-24.
66. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
67. Bittner, M., et al., *Molecular classification of cutaneous malignant melanoma by gene expression profiling.* Nature, 2000. **406**(6795): p. 536-40.
68. Su, A.I., et al., *Molecular classification of human carcinomas by use of gene expression signatures.* Cancer Res, 2001. **61**(20): p. 7388-93.
69. Rung, J., et al., *Building and analysing genome-wide gene disruption networks.* Bioinformatics, 2002. **18 Suppl 2**: p. S202-10.
70. Churchill, G.A., *Fundamentals of experimental design for cDNA microarrays.* Nat Genet, 2002. **32 Suppl**: p. 490-5.
71. Yang, Y.H. and T. Speed, *Design issues for cDNA microarray experiments.* Nat Rev Genet, 2002. **3**(8): p. 579-88.
72. Simon, R., M.D. Radmacher, and K. Dobbin, *Design of studies using DNA microarrays.* Genet Epidemiol, 2002. **23**(1): p. 21-36.
73. Kerr, M.K. and G.A. Churchill, *Statistical design and the analysis of gene expression microarray data.* Genet Res, 2001. **77**(2): p. 123-8.
74. Barrett, J.C. and E.S. Kawasaki, *Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression.* Drug Discov Today, 2003. **8**(3): p. 134-41.
75. Leung, Y.F. and D. Cavalieri, *Fundamentals of cDNA microarray data analysis.* Trends Genet, 2003. **19**(11): p. 649-59.
76. Hess, K.R., et al., *Microarrays: handling the deluge of data and extracting reliable information.* Trends Biotechnol, 2001. **19**(11): p. 463-8.
77. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 2006. **7**(1): p. 55-65.
78. Turk, R., et al., *Gene expression variation between mouse inbred strains.* BMC Genomics, 2004. **5**(1): p. 57.
79. Whitehead, A. and D.L. Crawford, *Variation in tissue-specific gene expression among natural populations.* Genome Biol, 2005. **6**(2): p. R13.
80. Hwang, D., W.A. Schmitt, and G. Stephanopoulos, *Determination of minimum sample size and discriminatory expression patterns in microarray data.* Bioinformatics, 2002. **18**(9): p. 1184-93.
81. Kerr, M.K., M. Martin, and G.A. Churchill, *Analysis of variance for gene expression microarray data.* J Comput Biol, 2000. **7**(6): p. 819-37.
82. Thygesen, H.H. and A.H. Zwinderman, *Comparing transformation methods for DNA microarray data.* BMC Bioinformatics, 2004. **5**: p. 77.

83. Holloway, A.J., et al., *Options available--from start to finish--for obtaining data from DNA microarrays II.* Nat Genet, 2002. **32 Suppl**: p. 481-9.
84. Kothapalli, R., et al., *Microarray results: how accurate are they?* BMC Bioinformatics, 2002. **3**: p. 22.
85. Blanchard, A.P., R.J. Kaiser, and L.E. Hood, *High-density oligonucleotide arrays.* Biosens. Bioelectron., 1996. **6/7**: p. 687-690.
86. Quackenbush, J., *Computational analysis of microarray data.* Nat Rev Genet, 2001. **2**(6): p. 418-27.
87. Hughes, T.R., et al., *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.* Nat Biotechnol, 2001. **19**(4): p. 342-7.
88. Kane, M.D., et al., *Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.* Nucleic Acids Res, 2000. **28**(22): p. 4552-7.
89. Kuo, W.P., et al., *Analysis of matched mRNA measurements from two different microarray technologies.* Bioinformatics, 2002. **18**(3): p. 405-12.
90. Yuen, T., et al., *Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays.* Nucleic Acids Res, 2002. **30**(10): p. e48.
91. Nimgaonkar, A., et al., *Reproducibility of gene expression across generations of Affymetrix microarrays.* BMC Bioinformatics, 2003. **4**: p. 27.
92. Jiang, H., et al., *Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes.* BMC Bioinformatics, 2004. **5**: p. 81.
93. Stillman, B.A. and J.L. Tonkinson, *Expression microarray hybridization kinetics depend on length of the immobilized DNA but are independent of immobilization substrate.* Anal Biochem, 2001. **295**(2): p. 149-57.
94. Southern, E., K. Mir, and M. Shchepinov, *Molecular interactions on microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 5-9.
95. t Hoen, P.A., et al., *Intensity-based analysis of two-colour microarrays enables efficient and flexible hybridization designs.* Nucleic Acids Res, 2004. **32**(4): p. e41.
96. Querec, T.D., et al., *A novel approach for increasing sensitivity and correcting saturation artifacts of radioactively labeled cDNA arrays.* Bioinformatics, 2004. **20**(12): p. 1955-61.
97. Schuchhardt, J., et al., *Normalization strategies for cDNA microarrays.* Nucleic Acids Res, 2000. **28**(10): p. E47.
98. Maskos, U. and E.M. Southern, *Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation.* Nucleic Acids Res, 1992. **20**(7): p. 1675-8.
99. Lyng, H., et al., *Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction.* BMC Genomics, 2004. **5**(1): p. 10.
100. Stoyanova, R., et al., *Normalization of single-channel DNA array data by principal component analysis.* Bioinformatics, 2004. **20**(11): p. 1772-84.
101. Eisen, M.B. and P.O. Brown, *DNA arrays for analysis of gene expression.* Methods Enzymol, 1999. **303**: p. 179-205.
102. Patriotis, P.C., et al., *ArrayExplorer, a program in Visual Basic for robust and accurate filter cDNA array analysis.* Biotechniques, 2001. **31**(4): p. 862, 864, 866-8, 870, 872.
103. Lawrence, N.D., et al., *Reducing the variability in cDNA microarray image processing by Bayesian inference.* Bioinformatics, 2004. **20**(4): p. 518-26.
104. Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 2002. **30**(4): p. e15.
105. Chudin, E., et al., *Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays.* Genome Biol, 2002. **3**(1): p. RESEARCH0005.
106. Sasik, R., E. Calvo, and J. Corbeil, *Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model.* Bioinformatics, 2002. **18**(12): p. 1633-40.
107. Kooperberg, C., et al., *Improved background correction for spotted DNA microarrays.* J Comput Biol, 2002. **9**(1): p. 55-66.
108. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.
109. Selvey, S., et al., *Beta-actin--an unsuitable internal control for RT-PCR.* Mol Cell Probes, 2001. **15**(5): p. 307-11.
110. Lee, P.D., et al., *Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies.* Genome Res, 2002. **12**(2): p. 292-7.

111. Kepler, T.B., L. Crosby, and K.T. Morgan, *Normalization and analysis of DNA microarray data by self-consistency and local regression.* Genome Biol, 2002. **3**(7): p. RESEARCH0037.
112. Zhao, Y., M.C. Li, and R. Simon, *An adaptive method for cDNA microarray normalization.* BMC Bioinformatics, 2005. **6**(1): p. 28.
113. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.
114. Chen, J.J., et al., *Analysis of variance components in gene expression data.* Bioinformatics, 2004. **20**(9): p. 1436-46.
115. Hoyle, D.C., et al., *Making sense of microarray data distributions.* Bioinformatics, 2002. **18**(4): p. 576-84.
116. Sapir, M. and G.A. Churchill, *Estimating the posterior probability of differential gene expression from microarray data*, T.J. Laboratory, Editor. 2000.
117. Tschentscher, F., et al., *Tumor classification based on gene expression profiling shows that uveal melanomas with and without monosomy 3 represent two distinct entities.* Cancer Res, 2003. **63**(10): p. 2578-84.
118. Chiappetta, P., M.C. Roubaud, and B. Torresani, *Blind source separation and the analysis of microarray data.* J Comput Biol, 2004. **11**(6): p. 1090-109.
119. Bohen, S.P., et al., *Variation in gene expression patterns in follicular lymphoma and the response to rituximab.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1926-30.
120. Oba, S., et al., *A Bayesian missing value estimation method for gene expression profile data.* Bioinformatics, 2003. **19**(16): p. 2088-96.
121. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520-5.
122. Kaski, S., et al., *Trustworthiness and metrics in visualizing similarity of gene expression.* BMC Bioinformatics, 2003. **4**: p. 48.
123. Zhou, X., X. Wang, and E.R. Dougherty, *Missing-value estimation using linear and non-linear regression with Bayesian gene selection.* Bioinformatics, 2003. **19**(17): p. 2302-7.
124. Hastie, T., et al., *Imputing missing data for gene expression arrays*. 1999, Department of Statistics Stanford University.
125. Kim, K.Y., B.J. Kim, and G.S. Yi, *Reuse of imputed data in microarray analysis increases imputation efficiency.* BMC Bioinformatics, 2004. **5**: p. 160.
126. Riva, A., et al., *The difficult interpretation of transcriptome data: the case of the GATC regulatory network.* Comput Biol Chem, 2004. **28**(2): p. 109-18.
127. Hirai, M.Y., et al., *Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana.* Proc Natl Acad Sci U S A, 2004. **101**(27): p. 10205-10.
128. Leung, Y.F., *Unravelling the mystery of microarray data analysis.* Trends Biotechnol, 2002. **20**(9): p. 366-8.
129. Sontag, E., A. Kiyatkin, and B.N. Kholodenko, *Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data.* Bioinformatics, 2004. **20**(12): p. 1877-86.
130. Davis, S.J. and A.J. Millar, *Watching the hands of the Arabidopsis biological clock.* Genome Biol, 2001. **2**(3): p. REVIEWS1008.
131. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**(5338): p. 680-6.
132. Newton, M.A., et al., *On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.* J Comput Biol, 2001. **8**(1): p. 37-52.
133. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
134. Benjamini, Y., et al., *Controlling the false discovery rate in behavior genetics research.* Behav Brain Res, 2001. **125**(1-2): p. 279-84.
135. Yeung, K.Y., M. Medvedovic, and R.E. Bumgarner, *From co-expression to co-regulation: how many microarray experiments do we need?* Genome Biol, 2004. **5**(7): p. R48.
136. Somorjai, R.L., B. Dolenko, and R. Baumgartner, *Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions.* Bioinformatics, 2003. **19**(12): p. 1484-91.
137. Martoglio, A.M., et al., *A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer.* Bioinformatics, 2002. **18**(12): p. 1617-24.

138. Alter, O., P.O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 10101-6.
139. Willenbrock, H. and D.W. Ussery, *Chromatin architecture and gene expression in Escherichia coli.* Genome Biol, 2004. **5**(12): p. 252.
140. Lovett, S.T. and A.M. Segall, *New views of the bacterial chromosome.* EMBO Rep, 2004. **5**(9): p. 860-4.
141. Niki, H., Y. Yamaichi, and S. Hiraga, *Dynamic organization of chromosomal DNA in Escherichia coli.* Genes Dev, 2000. **14**(2): p. 212-23.
142. Zimmerman, S.B., *Underlying regularity in the shapes of nucleoids of Escherichia coli: implications for nucleoid organization and partition.* J Struct Biol, 2003. **142**(2): p. 256-65.
143. Valens, M., et al., *Macrodomain organization of the Escherichia coli chromosome.* Embo J, 2004. **23**(21): p. 4330-41.
144. Audit, B. and C.A. Ouzounis, *From genes to genomes: universal scale-invariant properties of microbial chromosome organisation.* J Mol Biol, 2003. **332**(3): p. 617-33.

Résumé

Les mesures des niveaux d'expression de tous les gènes d'un génome requièrent une analyse statistique afin d'obtenir des conclusions fiables. Les biologistes ont du mal à faire un choix dans la foule de méthodes existantes et les comparaisons actuellement employées reposent sur des critères lacunaires ou non biologiques.

L'organisation du génome bactérien permet la définition d'un critère de comparaison à pertinence biologique indépendante de la problématique : les opérons. Ce critère a permis de procéder à une évaluation critique de méthodes classiques. Par ailleurs les méta-analyses de transcriptome sont en train de se développer malgré les biais inhérents à cette technologie. Elles ouvrent la possibilité d'étudier de nouveaux champs en biologie comme l'organisation chromosomique de l'expression des gènes. L'étude de trois bactéries, *B. subtilis, E. coli et S. meliloti* a révélé des corrélations d'expression à longue distance (environ 600kb) quel que soit le gène étudié.

Abstract

The analysis of transcriptomic data requires statistical methods to provide reliable conclusions. Amongst the huge amount of available methods, biologists may have difficulties to choose the most appropriate one for their needs. The existing criterions to compare different methods are either incomplete or use criteria that are not biologically relevant.

The organisation of bacterial genomes offers a biologically relevant criterion to compare the methods independently of the goal of the experiment: the operons. We have developed a protocol based on this criterion and compared some classical methods: PCA, ICA, t-test and ANOVA.

Furthermore, meta-analysis of transcriptome data is currently developed. These meta-analyses allow the study of new biological fields such as the chromosomal organisation of gene expression. We have analysed three bacteria, *B. subtilis, E. coli* and *S. meliloti* and have revealed long-range correlations of expression in all organisms, whatever the gene studied.