Les données d'expression

Nancie Reymond, Hubert Charles, Sophie Rome, Jacques Marti

Chapitre 2

Les données d'expression

2.1. Introduction

L'avènement de la transcriptomique a fait émerger une classe de données entièrement nouvelle en biologie. Les progrès fulgurants de ces dix dernières années réalisés dans les domaines de la micro-informatique et de la microfluidique, associés au génie génétique (clonage de l'ADN, PCR [1] et séquençage), ont permis un changement d'échelle dans la quantité de données acquises au cours d'une même expérience, ainsi qu'une miniaturisation de la plupart des appareillages. En effet, il est maintenant possible d'obtenir des estimations simultanées des niveaux d'expression de plusieurs milliers de gènes d'un organisme, d'un tissu ou même d'une seule cellule grâce à une expérience unique [BRA 00, GER 02].

L'analyse du transcriptome, c'est-à-dire le dénombrement absolu ou relatif des différents ARNm exprimés dans une cellule (ou dans un tissu) et dans une condition expérimentale donnée, trouve de nombreuses applications dans des domaines aussi divers que la médecine, la biologie fondamentale ou la microbiologie. Ainsi, l'utilisation des techniques d'analyse du transcriptome a permis la caractérisation de gènes impliqués dans certains cancers, la détection de mutations uniques (SNP) dans le génome humain, la détection de bactéries pathogènes ou encore le décryptage de réseaux de régulation génétique pour la levure ou chez diverses bactéries. Ces différentes applications sont plus particulièrement détaillées dans le chapitre 1.

---

1. *Polymerase chain reaction* : cette méthode permet d'amplifier *in vitro* une séquence d'ADN ou d'ARN connue à partir d'une faible quantité d'acide nucléique au départ.

Chapitre rédigé par Nancie REYMOND, Hubert CHARLES, Sophie ROME, Jacques MARTI.

On peut considérer qu'il existe trois types de méthodes d'analyse globale permettant d'obtenir des données d'expression : les méthodes basées sur l'utilisation de la PCR (expression différentielle et PCR soustractive), les méthodes basées sur le séquençage (EST [2], SAGE [3] et MPSS [4]) et celles basées sur le principe de l'hybridation moléculaire (puces à ADN). Les techniques SAGE et puces à ADN sont actuellement les stratégies dominantes de l'analyse du transcriptome car elles permettent de générer les plus importants volumes de données. Ces deux techniques seront plus particulièrement décrites dans ce chapitre.

D'une façon générale, la technique d'expression différentielle (*differential display reverse transcription polymerase chain reaction*, DDRT-PCR) consiste à comparer directement les populations d'ARN messagers (ARNm) de différents tissus. Ces ARNm sont copiés en ADN complémentaires (ADNc) grâce à une transcriptase inverse (RT) puis amplifiés par PCR. Les produits PCR sont séparés par électrophorèse et les profils électrophorétiques sont comparés visuellement de façon à détecter les gènes différentiellement exprimés dans les échantillons. Cette technique, décrite initialement en 1992 par Liang et Pardee [LIA 92], a été très améliorée, mais le principe reste le même.

La technique de PCR soustractive, comme son nom l'indique, est basée sur une étape de soustraction des deux populations d'ADNc que l'on souhaite comparer. Cette soustraction est effectuée par hybridation. Les ADNc communs aux deux populations s'apparient entre eux lorsqu'ils sont mélangés. En utilisant la technique de PCR associée à des jeux d'adaptateurs spécifiques, seuls les ADNc spécifiques de l'une des deux conditions sont ensuite amplifiés par PCR. On se référera au travail de Welsh et McClelland [WEL 90] pour une description détaillée de cette technique.

Parmi les techniques d'analyse du transcriptome basées sur le séquençage, la collection d'EST [2] a été pionnière. Cette technique est très simple dans son principe, puisqu'elle consiste à cloner les deux populations d'ADNc que l'on souhaite comparer pour réaliser des banques d'ADNc. Les clones sont alors séquencés (entièrement ou partiellement) sans sélection. Les séquences obtenues sont appelées EST et sont des marqueurs de séquences exprimées. Le décompte des différentes copies des gènes identifiés est une estimation de leur niveau d'expression. Actuellement, compte tenu du coût associé à la création de banques d'EST utilisables en analyse quantitative, la méthode SAGE qui utilise également la technique de séquençage mais sur de très courts fragments spécifiques d'ARNm appelés « tags » est utilisée préférentiellement [VEL 95]. Cette méthode sera décrite de façon plus complète dans la section suivante.

---

2. *Expressed sequence tag*.
3. *Serial analysis of gene expression*.
4. *Massively parallel signature sequencing*.

Enfin, une méthode très prometteuse mais encore peu répandue vient de voir le jour. Il s'agit de la technologie MPSS développée par la société LYNX [5]. Cette technique est basée sur l'utilisation de collections de microbilles sur lesquelles sont greffés des « antitags » (technique Megaclone^{TM}). Ces « antitags » sont de très courts oligonucléotides qui agissent comme des anticorps permettant de fixer sélectivement chacune des copies (ou fragments de copie) d'ADNc de la collection. Une technologie de séquençage colorimétrique utilisant également la technologie des microbilles a été développée (technique MPSS). Cette technologie permet d'identifier simultanément, au cours d'une même expérience, plusieurs millions de copies d'ARNm [BRE 00]. Il semble que cette nouvelle technologie permette un véritable changement d'échelle par rapport à l'analyse SAGE (dans un rapport de 1 pour 1000 en nombre de tags identifiés). A l'heure actuelle, elle reste néanmoins très difficilement abordable en raison du coût élevé.

Enfin, les méthodes utilisant l'hybridation sont toutes basées sur le principe de Southern [SOU 74] : deux fragments d'acides nucléiques peuvent s'associer et se dissocier (de façon réversible) sous l'action de la chaleur et de la concentration saline du milieu. La stabilité du duplex formé dépend à la fois de la similarité entre les deux séquences et de leur composition en bases. Dans le cas d'une analyse à large échelle, il est possible de réaliser cette hybridation sur une phase solide, comme une lame de verre par exemple, ce qui permet de travailler simultanément avec un grand nombre de sondes [6] dont les positions sont bien déterminées. L'utilisation de microtechnologies permet d'analyser simultanément plusieurs milliers de gènes sur des supports de l'ordre du cm^2. Historiquement, la désignation de « puce à ADN » (*DNA chip*) était restreinte à des lames de verre sur lesquelles les sondes (fragments de quinze à vingt bases) étaient synthétisées *in situ* à très haute densité (100 000 sondes/cm^2). Ces puces sont commercialisées par la société Affymetrix [7]. Le terme de « *microarray* » était réservé aux supports en nylon, avec une moindre densité de sondes (1 000 à 10 000 par cm^2) déposées par un robot (*spotter*) et une révélation radioactive. Avec l'évolution des technologies (utilisation de la fluorescence sur verre ou de la synthèse *in situ* d'oligonucléotides longs), le terme de puce à ADN est utilisé actuellement de façon générique. Cette technique est actuellement en plein essor [CAS 98, GER 02, LEM 98]. Son principe est décrit dans la section suivante.

2.2. Acquisition des données d'expression

Dans ce chapitre, nous proposons une description du principe des techniques SAGE et puces à ADN qui sont les deux méthodes d'analyse du transcriptome les

5. Http://www.lynxgen.com.
6. ADN synthétique complémentaire du gène que l'on souhaite étudier.
7. Http://www.affymetrix.com.

plus couramment utilisées. Plutôt que de donner une vision exhaustive des procédures très complexes mises en œuvre, nous essaierons de nous focaliser uniquement sur les aspects technologiques susceptibles d'influencer la nature des données collectées (en termes de variabilité, reproductibilité, sensibilité, etc.).

2.2.1. *Le SAGE ou analyse sérielle de l'expression génique*

La méthode SAGE [VEL 95] est une évolution logique des approches par analyse d'EST, qu'elle améliore en particulier sur deux points :

– la taille des séquences analysées est réduite au minimum. On peut en effet considérer, dans le cas d'une espèce dont le génome commence à être bien annoté, qu'une séquence d'ADNc de dix paires de bases, copiée sur un site précis d'un ARNm, constitue une signature (tag) suffisante pour l'identifier ;

– les tags sont d'abord amplifiés par une méthode originale, qui conserve les proportions relatives des ARNm initiaux, puis liés en concatémères pour être séquencés en série.

On peut ainsi analyser un grand nombre de tags, et non seulement identifier la majorité des ARNm d'un extrait cellulaire mais également mesurer, d'après la fréquence de chaque tag, le niveau d'expression des gènes correspondants.

Pratiquement, les ARNm sont d'abord capturés, par leur extrémité 3′ polyadénylée, sur des billes magnétiques et convertis en ADNc directement sur ce support. Une première digestion enzymatique, éliminant environ 80 % de la masse d'ADNc, laisse fixé aux billes le fragment 3′-terminal de chaque molécule. Tous les fragments issus de la digestion portent alors la même séquence en 5′, habituellement CATG si l'on utilise l'enzyme *Nla III* (certains laboratoires utilisent l'enzyme *Sau 3A1* reconnaissant les sites GATC). Les sites de quatre bases de ces deux enzymes sont très fréquents (1/256) et pratiquement tous les ADNc en possèdent au moins un. Ainsi, seuls les rares transcrits dépourvus de site de coupure échapperont à l'analyse (voir figure 2.1).

L'étape suivante consiste à greffer un oligonucléotide de synthèse (adaptateur) sur ces terminaisons CATG de façon à créer le site de fixation de *Bsm f1*, enzyme qui génère les tags en coupant chaque ADNc quatorze bases au-delà de ce site. Les fragments, ainsi libérés des billes magnétiques, portent en 5′ la séquence de l'adaptateur et diffèrent en 3′ par les dix bases spécifiques de chaque tag. Une variante appelée « Long SAGE » utilise l'enzyme *Mme 1* pour générer des tags plus longs.

Un apport original de la méthode SAGE est de réaliser l'opération en double, sur deux parties aliquotes du même échantillon, avec deux adaptateurs différents. On obtient ainsi deux populations de tags présentant la même distribution statistique. Une ligase permet de les réunir en « ditags » par leur extrémité 3′, les parties synthétiques étant conçues pour ne pas réagir. Les ditags sont amplifiés par PCR avec deux

amorces, chacune spécifique de l'un des deux adaptateurs. Cette amplification permet de construire une banque SAGE à partir d'un très petit nombre de cellules.

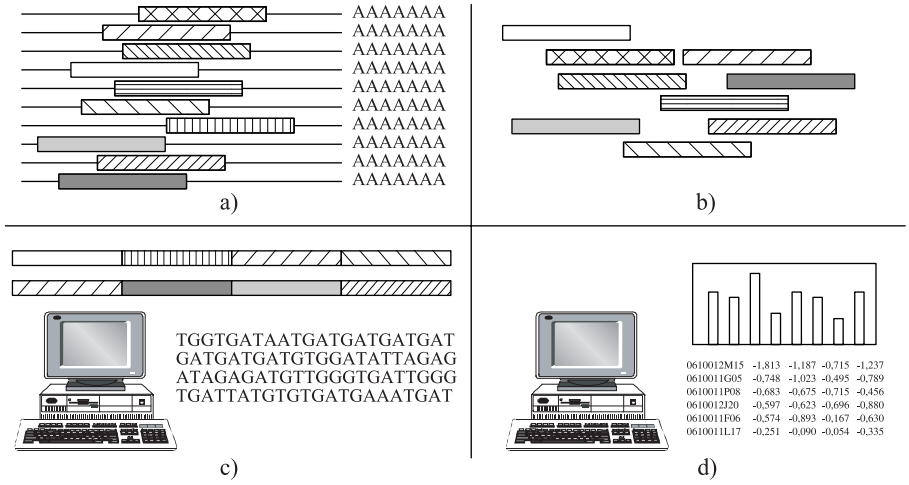

Figure 2.1. *Principe de la technique SAGE. (a) Les ARNm sont capturés par leur extrémité 3′ polyadénylée. (b) Les tags sont libérés par coupure enzymatique. (c) Après concaténation des tags, les concatémères sont séquencés en série. (d) Les différents tags sont enfin comptabilisés et identifiés grâce aux bases de données.*

Les ditags sont ensuite libérés des adaptateurs par l'enzyme déjà utilisée pour générer leurs sites d'ancrages (CATG), puis liés en concatémères. Leur assemblage aléatoire fait de chaque concatémère une molécule unique, qui est amplifiée par clonage bactérien avant d'être séquencée. Cette étape est critique, car la taille des concatémères conditionne la quantité d'information portée par chaque clone, tous séquencés pour un même coût unitaire. La décision d'arrêter l'analyse d'une banque SAGE à partir d'un certain nombre de tags (habituellement de 20 000 à 60 000) dépend souvent d'un compromis entre le coût et le but recherché. Les données brutes se présentent sous forme d'un fichier de séquences.

Plusieurs logiciels sont disponibles pour reconnaître la séquence (CATG) commençant chaque tag, enregistrer la chaîne des dix caractères suivants et déterminer son nombre d'occurrences. Comme les tags ont été liés en orientation inverse dans les ditags, il faut lire chacun des deux brins de l'ADN séquencé. Le résultat final est une liste donnant la séquence et le nombre d'occurrences de chaque tag. Pour identifier les ARNm correspondants, il est pratique de dresser, indépendamment de l'expérience, la liste des tags virtuels susceptibles d'être observés d'après les connaissances du génome de l'espèce considérée. Avec les outils usuels d'un système de gestion de base de données, on peut alors faire correspondre tags virtuels et expérimentaux, du

moins pour les gènes déjà connus. Les mesures résultant d'un simple dénombrement, les données d'expériences indépendantes peuvent être comparées directement et le niveau d'expression d'un même tag peut être comparé sur l'ensemble des banques SAGE publiées.

2.2.2. *Les puces à ADN*

L'idée conceptuelle de la puce à ADN est très simple. Il s'agit de greffer sur une surface de quelques centimètres carrés des fragments synthétiques d'ADN (les sondes) représentatifs de chacun des gènes que l'on souhaite étudier et espacés de quelques micromètres. Ce microdispositif est ensuite mis au contact des acides nucléiques à analyser, les ARNm ou les ADNc (appelés cibles) qui ont été préalablement couplés à un marqueur fluorescent ou radioactif. Ce contact entre les cibles et les sondes conduit à la formation de duplex que l'on peut qualifier, par leurs coordonnées, et quantifier grâce à la lecture des signaux radioactifs ou fluorescents (voir figure 2.2).

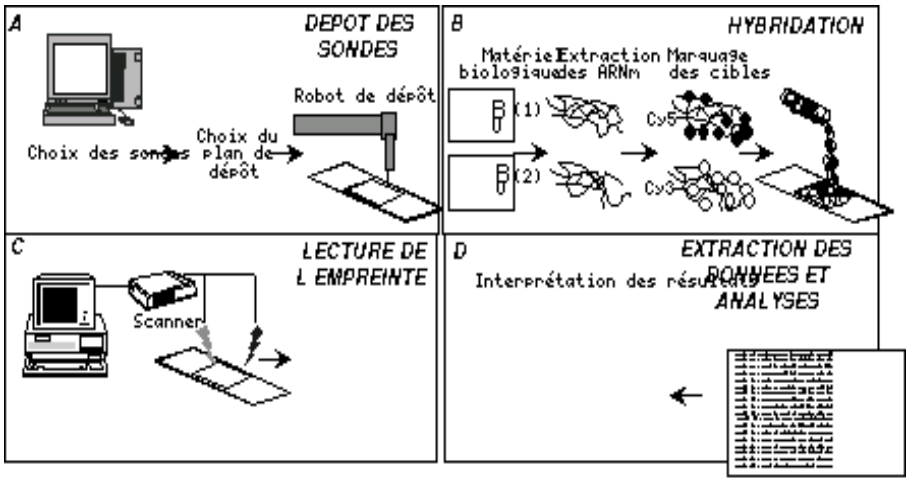

Figure 2.2. *Principe de la technique des puces à ADN. (a) Les séquences des sondes sont déterminées de façon à optimiser leur spécificité et leur sensibilité. Les sondes synthétisées sont déposées selon un plan défini par un robot sur la surface de la lame. (b) Les ARNm sont extraits des échantillons biologiques à comparer, marqués grâce à deux fluorochromes différents puis mélangés avant l'hybridation. (c) La lecture des lames est réalisée grâce à un scanner (microscope à fluorescence) couplé à un photomultiplicateur (PMT). (d) L'image est alors analysée de façon à quantifier le signal. Les données seront ensuite normalisées, analysées et interprétées.*

Il existe une diversité de supports pour la fabrication des puces à ADN. Les premières puces à ADN développées (*macroarrays*) utilisaient un support nylon. Ce support possède de très bonnes propriétés pour la fixation de la sonde ; par contre, la

surface ne permet pas un dépôt à très forte densité de sondes. La technologie nylon reste néanmoins utilisée et très performante [CAO 02]. Les supports neutres les plus couramment utilisés sont le verre et le plastique, mais il existe un certain nombre de supports dits « actifs », comme le silicium. Ces supports voient leurs propriétés modifiées (par exemple leur conductivité) par la formation du duplex au cours de l'hybridation. Ils participent ainsi directement à la détection du signal. Ce type de technologie prometteur est en plein développement [BEL 97].

Les sondes greffées ou synthétisées peuvent être de différentes tailles. Les puces à sondes courtes (15-20 pb) sont caractérisées par une spécificité très forte (une seule base de différence entre la cible et la sonde suffit, en théorie, à interdire l'hybridation) et une sensibilité faible. Elles sont principalement utilisées pour le génotypage, les cibles étant généralement de l'ADN génomique. Un des verrous technologiques de ces puces reste la possibilité d'effectuer des mesures quantitatives. La technologie Affymetrix utilise également des sondes courtes pour l'analyse d'expression de gènes mais en multipliant le nombre de sondes par gène (voir ci-dessous). Les puces à ADN à sondes oligonucléotidiques moyennes (30-70 pb) sont utilisées principalement pour l'analyse du transcriptome. Elles sont caractérisées par une bonne spécificité et une bonne sensibilité. Enfin, il est possible de greffer sur les lames des fragments d'ADN entiers issus d'amplification par PCR (100 à 500 pb). Ces puces à ADN présentent une sensibilité maximale mais sont, en théorie, moins bonnes en termes de spécificité. La détermination de la séquence de la sonde nécessite une analyse bioinformatique de façon à optimiser à la fois les paramètres de spécificité (la sonde ne doit reconnaître qu'un seul gène) et les paramètres thermodynamiques (la stabilité du duplex). Des logiciels d'optimisation de sondes ont été développés et sont disponibles pour la communauté académique [REY 04, ROU 02].

Les sondes peuvent être produites avant l'étape d'accrochage ou directement sur le support. Lorsque les sondes sont des oligonucléotides de synthèse ou des fragments d'ADN, ceux-ci sont déposés grâce à un robot sur la lame. Le type d'accrochage peut être de type électrostatique ou covalent. Dans ce dernier cas, l'oligonucléotide est modifié à l'une de ses extrémités et porte un groupement réactif (une amine par exemple). La surface de la lame est traitée de façon complémentaire pour présenter des groupements actifs (comme des aldéhydes) capables de générer des liaisons covalentes avec l'oligonucléotide modifié. La qualité du signal détecté sur la puce dépend de la qualité du support, du dépôt par le robot et des conditions d'hygrométrie dans lesquelles se déroule la fixation. Une autre façon de fabriquer des puces à ADN consiste à synthétiser les sondes directement sur le support (synthèse *in situ*). Deux grands types de techniques sont actuellement développés : la microsynthèse et la synthèse photolithographique (procédé Affymetrix). Pour une description détaillée de ces techniques, on pourra se référer aux travaux [HEL 02, HUG 01].

Enfin, le niveau d'expression de chacun des gènes est représenté sur la puce par un à quelques spots de sondes pour ce qui concerne les puces de faible et moyenne

densité. Après traitement de l'image, le signal analysé est, en général, la moyenne (ou la médiane) des pixels composant chaque plot, moyenne à laquelle on retranche souvent une valeur de bruit de fond estimée dans la zone périphérique du plot (voir paragraphe 2.3.3.3). Des répétitions (ou réplicats) de ces sondes sont réparties sur la surface de lame de façon à accéder à une part de la variabilité locale du signal. Les puces à haute densité, commercialisées par la société Affymetrix, ont une structure très particulière. Pour chaque gène, une série de dix à vingt sondes, réparties sur toute la séquence du gène, est représentée sur la lame. A chacune de ces sondes PM (*perfect match*) est associée une sonde MM (*mismatch*) dont la séquence est identique à la séquence PM mais avec une mutation ponctuelle située en position centrale. La sonde MM permet de quantifier la part du signal aspécifique (bruit de fond) associé à la sonde PM. Le calcul du niveau d'expression d'un gène est relativement complexe, mais peut être considéré, en première approximation, comme une moyenne pondérée des différences (PM-MM) de chaque paire de sondes associées à ce gène. Pour plus de détails, on se référera au travail [CHU 02].

2.3. Caractérisation des données d'expression

L'acquisition de données d'expression globales – que ce soit par la méthode SAGE ou grâce aux puces à ADN – inclut de nombreuses étapes plus ou moins complexes et contrôlées. Chacune de ces étapes a une source de variabilité propre qui doit être prise en considération pour une meilleure normalisation des résultats. L'analyse statistique ou informatique vise à tester des hypothèses biologiques en comparant des niveaux d'expression de gènes impliqués dans des processus biologiques. Pour cela, il faut que l'estimation du niveau d'expression – et sa variabilité associée – soit non biaisée. En d'autres termes, il faut réussir à caractériser et à modéliser la part de variation liée à l'expérimentation.

2.3.1. *Complexité des données d'expression*

Les données d'expression représentent des volumes importants. Toutes les techniques d'analyse du transcriptome permettent actuellement d'estimer simultanément les niveaux d'expression de plusieurs milliers de gènes dans plusieurs conditions expérimentales. Faire face au stockage et à l'analyse (tests multiples en statistique par exemple) de tels volumes de données représente une difficulté notable. De plus, le nombre de conditions expérimentales (par exemple le nombre d'échantillons prélevés sur un ensemble de patients) est bien souvent faible par rapport aux nombres de gènes analysés. Il existe donc une dissymétrie importante du tableau de données, c'est-à-dire une redondance extrême des régresseurs si l'on cherche à caractériser les conditions expérimentales par les gènes, et une sous-paramétrisation si l'on souhaite adopter la démarche inverse.

Les données d'expression sont des mesures relatives. Dans le cas des puces à ADN, en première approximation, on peut considérer que pour un gène donné l'intensité du signal est proportionnelle à l'abondance de l'ARNm correspondant dans la solution de cibles. Ce coefficient de proportionnalité est lié à l'affinité de la sonde pour sa cible qui peut être mesurée en partie par sa température de fusion [8] (T_m ou *melting temperature*). Comme les hybridations pour toutes les sondes de la même puce sont réalisées simultanément (à la même température), il paraît indispensable que toutes les sondes aient des T_m identiques pour pouvoir quantifier de façon absolue les abondances de chaque ARNm du tissu étudié. De plus, lorsque la fluorescence est utilisée, la mesure est amplifiée par un photomultiplicateur (PMT). Cette détection ne permettant pas l'obtention de valeurs absolues de la mesure, une analyse en double marquage est le plus souvent réalisée (voir figure 2.2). Les rapports des signaux analysés constituent alors une mesure relative entre les deux conditions testées. Pour les puces à haute densité (Affymetrix), et dans certains cas d'analyses utilisant la radioactivité, la mesure relative est calculée en effectuant les rapports de signaux issus de deux puces distinctes hybridées chacune avec les ARN messagers des deux types de tissus. Cette calibration sera évoquée dans le chapitre 3. Dans le cas de l'analyse SAGE, on peut considérer, à première vue, que les données sont de nature plus absolue puisque l'on dénombre directement les transcrits isolés à l'aide de leurs « tags » correspondants. Il faut néanmoins conserver à l'esprit que le « niveau zéro » n'existe pas dans cette technologie. En effet, un effort supplémentaire de séquençage offrira toujours une probabilité supplémentaire aux gènes considérés « non exprimés » d'apparaître dans la liste des gènes « exprimés ». La mesure est donc relative, mais contrairement au cas des puces à ADN, ce sont les niveaux d'expression des gènes qui sont relatifs au sein d'une même expérience et non plus chaque niveau d'expression entre les deux conditions expérimentales. Les contraintes de calibrage sont donc moins élevées avec cette technologie, à condition de comparer des expérimentations de tailles identiques.

Les données d'expression ne sont pas distribuées normalement. On peut considérer que dans la plupart des organismes cellulaires (des bactéries les plus simples aux eucaryotes les plus organisés), à un instant donné, la grande majorité des gènes s'exprime à un niveau très faible (voir figure 2.3). Le niveau d'expression d'un gène n'est pas lié à l'importance de sa fonction. Les facteurs de transcription, par exemple, sont exprimés à des niveaux très faibles alors que certains très ubiquitaires sont à l'origine de changements physiologiques majeurs dans la cellule. La distribution des niveaux d'expression est donc systématiquement dissymétrique et très étirée vers les fortes valeurs d'expression (figure 2.3).

Les données d'expression sont caractérisées par la non-indépendance des gènes. En effet, la plupart des mécanismes biologiques sont régulés au niveau transcriptionnel

8. Température à laquelle 50 % des sondes sont associées à leur cible.

par des cascades d'activateurs ou d'inhibiteurs, certains facteurs de transcription étant capables d'activer ou d'inhiber l'expression de plusieurs dizaines de gènes à la fois. Cette dépendance est un facteur très important à prendre en considération pour les analyses statistiques.

La dernière caractéristique des données d'expression est l'hétérogénéité des connaissances qui leur sont associées (voir section 2.4), ceci étant en partie lié à la mise à jour continuelle de l'annotation des génomes. Pour ne donner qu'un seul exemple, la notion de « gène » reste extrêmement délicate à définir en raison, entre autres, des phénomènes d'épissages alternatifs et de l'annotation incomplète des génomes. Est-ce que le « tag » ou la « sonde » utilisés seront caractéristiques du seul, de quelques-uns ou de tous les transcrits d'un même gène et connaît-on tous les transcrits de ce gène ?

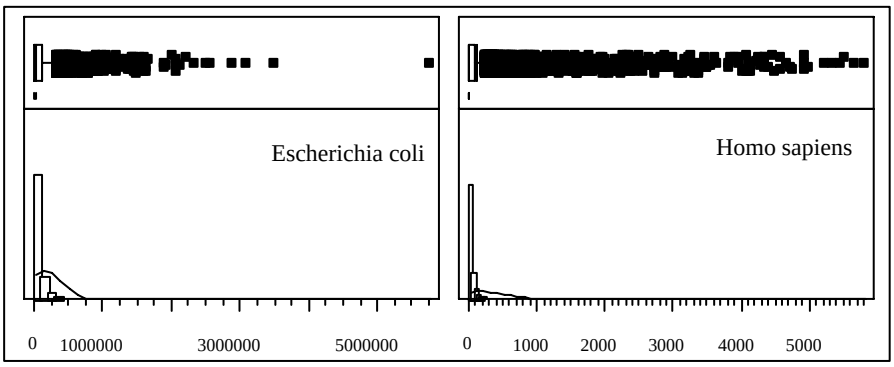

Figure 2.3. *Distribution des niveaux d'expression (unité relative de fluorescence en abscisse) chez la bactérie* Escherichia coli *(5 522 gènes) et chez l'homme (12 632 gènes). Les diagrammes en « box-plots » montrent l'étalement de la distribution vers les fortes valeurs (faible nombre de gènes fortement exprimés) et la courbe représente la distribution normale théorique. Les données sont issues de la base d'expression GEO (www.ncbi.nih.gov/geo, numéros d'accession : GSE33 et GSE516).*

2.3.2. *Variabilité biologique des données d'expression*

Le matériel biologique est bien souvent un matériel hétérogène. Par exemple, une tumeur est constituée de nombreuses cellules très différentes ; de même, une population bactérienne est composée de cellules dans des états physiologiques très variables. Cependant, travailler sur une cellule unique, en plus de la miniaturisation nécessaire, ne constitue pas une solution alternative, car les réponses individuelles de chacune des cellules sont également très variables [BLA 03]. L'idéal serait donc de travailler sur

une population de cellules homogènes et synchronisées. La pratique est malheureusement souvent très éloignée de cet idéal.

L'ARNm est une molécule relativement instable dans la cellule. Les demi-vies de ces molécules sont extrêmement variables [ROS 96]. La notion de transcriptome est, par nature, une notion dynamique, et des études cinétiques devront souvent être réalisées. De plus, les ARNm ne seront pas traduits avec la même efficacité et des expériences complémentaires (en protéomique) seront parfois nécessaires pour interpréter les observations faites à un niveau phénotypique [WIN 03].

2.3.3. *Variabilité expérimentale des données d'expression*

Les techniques d'analyse du transcriptome impliquent des étapes multiples apportant chacune leur source d'erreur. Ces erreurs cumulées affectent la variabilité, la sensibilité et la spécificité de la mesure.

2.3.3.1. *Préparation du matériel biologique*

L'analyse de populations de cellules homogènes impose parfois de travailler sur des échantillons de taille très réduite. Certains modèles d'étude (bactéries non cultivables, prélèvements par microchirurgie, etc.) restreignent également de façon drastique les échantillons. Il est alors nécessaire de réaliser une étape d'amplification des ARNm avant l'analyse. Cette étape peut induire des biais très importants liés à des phénomènes d'amplifications sélectives. Pour une description d'une méthode d'amplification performante et adaptée à l'analyse du transcriptome, on pourra se reporter au travail [ISC 02].

L'extraction des ARN constitue en elle-même un stress cellulaire qui induit obligatoirement une réponse de l'organisme étudié. Cette réponse à l'extraction est, dans la plupart des cas, inscrite dans les données mesurées au final (par exemple par l'induction de gènes codant des protéines de choc thermique). Un protocole d'extraction efficace devra être mis en place (toujours sur le même exemple en incluant une congélation des tissus) et une planification expérimentale appropriée devra permettre une prise en considération de ce biais.

Même si les protocoles utilisés actuellement pour la préparation du matériel biologique sont très bien standardisés et montrent une bonne reproductibilité, la variabilité biologique que l'on détermine englobe toujours une part de variabilité expérimentale liée à l'extraction et au traitement de l'échantillon.

2.3.3.2. *Mesure des niveaux d'expression par la méthode SAGE*

Comme toute méthode analytique, l'approche SAGE demande une étude critique des résultats. Un aspect intéressant de cette méthode est qu'il est possible d'extraire

des informations sur la qualité des données à partir des fichiers de séquences eux-mêmes.

Un premier problème, qui n'est pas propre à la méthode SAGE, est le rendement de synthèse des ADNc, dépendant de l'efficacité de la transcriptase inverse qui varie d'un ARNm à l'autre. On notera cependant que les tags proviennent du site le plus proche de l'extrémité 3′, où le rendement est le plus élevé.

Il importe que l'enzyme coupant l'ADNc sur ce site ait elle-même une efficacité optimale. On peut le vérifier *a posteriori*, en recherchant les tags situés en 5′ du site attendu, du fait d'un défaut de coupure sur le site terminal. L'analyse des données montre que ce défaut de coupure existe à un taux faible, de l'ordre de 1 %.

La méthode SAGE implique une amplification par PCR qui pourrait fausser les mesures si chaque tag était amplifié séparément. Le fait de les associer en ditags offre plusieurs avantages. D'une part, tous sont amplifiés avec le même couple d'amorces. D'autre part, chaque ditag est une combinaison unique, ce qui réduit l'avantage sélectif lié à l'abondance spécifique d'un tag. Enfin, comme ils sont très courts, le rendement d'amplification est peu sensible à leur composition en bases. Il semble actuellement difficile d'imaginer un procédé respectant mieux les proportions initiales des ARNm.

Un avantage de la méthode SAGE est de pouvoir analyser de très petits échantillons. Cependant, amplifier à l'excès une petite population de tags peut générer une banque de complexité moindre que la population initiale d'ARNm. Il est possible de contrôler ce niveau de complexité en recherchant la présence de ditags identiques. En effet, du moins dans un extrait cellulaire contenant une grande diversité d'ARNm, la probabilité que deux tags différents s'associent plusieurs fois au sein du même ditag est très faible. L'expérience confirme que le taux de ditags identiques est plus élevé dans des banques SAGE réalisées à partir d'un échantillon très réduit, mais qu'il reste habituellement inférieur à 5 %.

Il serait tentant d'amplifier les clones bactériens après transfection, opération peu coûteuse mais qui risque de propager toujours les mêmes clones. Il est possible de s'en assurer en vérifiant l'absence de concatémères identiques. Si, pour compléter l'analyse, on souhaite séquencer de nouveaux clones, il est tout de même préférable de revenir à la préparation initiale de concatémères et de transfecter de nouvelles bactéries.

On peut évaluer le taux d'erreurs induit par l'analyse séquentielle en comptant le nombre de ditags apparemment deux fois trop longs, qui signalent la perte de sites CATG en cours d'analyse. Avec un taux de 1 % par position, on estime qu'un tag sur dix est erroné, estimation grossière qui ne tient pas compte de la répartition inégale des erreurs selon les séquences. La probabilité de reproduire plusieurs fois la même erreur étant très faible, ces tags erronés seront généralement uniques. Cependant, un

tag observé une seule fois dans une banque peut s'avérer fortement représenté dans d'autres. Il est donc avantageux de comparer chaque nouvelle banque à l'ensemble des banques disponibles. Sur 136 banques humaines publiées, soit six millions de tags, on recense 437 000 tags différents. Parmi eux, 43,5 % sont observés une seule fois, mais ces tags uniques ne représentent que 3,2 % du total des tags dénombrés. Par conséquent, bien qu'elles augmentent artificiellement le nombre de tags, les erreurs de séquences n'ont qu'une faible incidence sur le dénombrement des tags authentiques.

L'ambiguïté de l'assignation tag-gène peut avoir une incidence sur la mesure de certains gènes. Dans les génomes de vertébrés, le nombre des séquences de 14 pb (dont quatre communes et dix variables) dépasse le nombre de combinaisons statistiques et l'on ne peut assigner un tag à un seul locus. En fait, l'essentiel des ambiguïtés est levé si l'on considère que les tags proviennent des régions transcrites, qui ne représentent qu'une faible fraction de l'ADN génomique. La plupart des génomes étant partiellement annotés, les requêtes renvoient souvent à des fragments (EST). Les tags non identifiés sont conservés en attendant les progrès de l'annotation. Une partie provient de gènes connus, mais exprimant des transcrits encore inconnus, notamment en raison de la variabilité du site de terminaison : environ 40 % des gènes humains possèdent plus d'un site de polyadénylation. Un petit nombre de tags, correspondant à la séquence complémentaire inverse d'ARNm, sont susceptibles de signaler des transcrits antisens. Enfin, certains tags renvoient à de nombreux gènes partageant des similarités de séquence en 3′. Chez l'homme, ce sont principalement les séquences de type Alu qui introduisent cette ambiguïté : elle concerne environ 10 % de la totalité des ARNm d'un échantillon.

Ne dépendant pas d'un dispositif de mesure, la méthode SAGE ne demande pas de comparaison avec un étalon de référence. L'incertitude sur le nombre d'occurrences de chaque tag décroît avec le nombre de séquences analysées. On peut l'évaluer en assimilant les opérations (synthèse des concatémères et prélèvement des clones) à des tirages au hasard. Il est en effet difficile d'imaginer un biais dans le prélèvement des tags et, expérimentalement, on constate que la distribution du même tag dans des prélèvements successifs suit bien une loi binomiale. Dans une comparaison entre banques, on peut s'appuyer sur ce résultat pour calculer, sans qu'il soit nécessaire de normaliser les valeurs, la probabilité que les variations observées soient dues au hasard, ce qui permet corrélativement d'évaluer leur signification biologique.

2.3.3.3. *Mesure des niveaux d'expression par la méthode des puces à ADN*

La technologie des puces implique la fabrication de la lame support des sondes, une réaction d'hybridation et une lecture du signal. Ces trois étapes peuvent générer des biais et des sources de variabilités qu'il est nécessaire de discuter.

La fabrication de la lame support des sondes combine des problèmes de chimie (traitement de la surface de la lame et accrochage de l'ADN sonde) et de mécanique (dépôt des sondes par un robot sur la lame). Des disfonctionnements sur ces différentes étapes peuvent conduire à des irrégularités spatiales sur la lame qu'il faudra rechercher et éliminer dans les données. A titre d'exemple, une mauvaise calibration des aiguilles du robot de dépôt conduira à un « effet aiguille » dont il faut tenir compte lors de l'analyse des données. La répartition aléatoire (ou au moins uniforme) des sondes et des répétitions sur la lame prend alors tout son sens face à ces problèmes (ce point sera repris dans le chapitre 3).

Dans le cas des puces à ADN, le marquage des ARNm par accrochage d'une molécule fluorescente ou radioactive est réalisé au cours d'une réaction enzymatique impliquant la transcriptase inverse. L'enzyme utilisée est très sensible à de nombreux paramètres comme la taille de la molécule à incorporer, la composition en bases de la molécule d'ARNm, la température et la concentration saline du milieu. Cette sensibilité peut générer un marquage différentiel de chacun des gènes. Cependant, ces protocoles de marquage ont été relativement bien optimisés [HOE 03].

Effectuer en parallèle des milliers de réactions d'hybridation impliquant des séquences de compositions différentes et déterminer avec précision les hybridations spécifiques et aspécifiques représente une véritable difficulté technique. Les paramètres influençant la stabilité de l'hybridation ont été relativement bien étudiés [MAS 93] de façon à déterminer les conditions dans lesquelles la stabilité des duplex cibles/sondes est favorisée, notamment vis-à-vis des mésappariements (problème de l'hybridation aspécifique). Il existe cependant une variabilité importante de ces paramètres pour l'ensemble des sondes qui est difficilement maîtrisable (Tm, structures secondaires, etc.). De plus, ces études ont été réalisées pour des hybridations en milieu liquide et rien n'est connu concernant l'hybridation d'une cible avec sa sonde homologue fixée sur une matrice solide. Le contrôle de cette étape d'hybridation apparaît comme le verrou technologique le plus important pour s'affranchir de la mesure relative des niveaux d'expression avec cette technique.

La lecture du signal après hybridation constitue également une étape importante du protocole expérimental. Les cyanines fluorescentes Cy3 (vert) et Cy5 (rouge) sont très souvent utilisées pour marquer les cibles car elles possèdent un haut niveau d'émission photonique sous forme d'un pic étroit assurant une bonne sensibilité. Des quantités d'ARNm de l'ordre du picomolaire peuvent ainsi être détectées sur une puce pour un gène donné [BEL 97]. Cependant, des problèmes de stabilité des fluorophores et de *quenching* peuvent être observés entre les deux fluorochromes. Afin d'estimer et d'éliminer cet « effet fluorochrome », des marquages inverses des échantillons devront être prévus (analyse en *flip-flop* ou en *dye-swap*). Lors de la lecture des données, une saturation du signal, liée à une trop forte amplification par le tube photomultiplicateur, est souvent visible. La saturation se traduit par une distribution tronquée vers les

fortes valeurs dans le profil d'expression. Les spots saturés doivent donc être éliminés des analyses. Enfin, des problèmes de diffusion/réfraction se traduisant par une autocorrélation spatiale des spots et une non-linéarité du signal peuvent être détectés et éventuellement corrigés dans les données.

Le dernier problème lié à la mesure du niveau d'expression des gènes par la technique des puces à ADN concerne la mesure du bruit de fond (niveau zéro d'expression). Ce problème est d'autant plus exacerbé que, d'une part, la majorité des gènes est faiblement exprimée et que, d'autre part, la détection des rapports de fluorescence correspondant aux faibles niveaux est entachée d'une grande variabilité liée à la mesure relative (division par zéro). Certains auteurs préconisent de ne pas estimer ni retrancher de valeur de bruit de fond, lorsque l'analyse réalisée ensuite incorpore cette estimation (cas d'une analyse de variance par exemple). D'autres auteurs préconisent, à l'inverse, une filtration drastique des données de façon à éliminer cette queue de distribution pour ramener la distribution des niveaux d'expression vers une distribution plus normale. Ces points de vue sont abordés au chapitre 3.

2.4. Vers un formalisme de la représentation des données d'expression

L'effort très important réalisé par la communauté scientifique a permis à la plupart des laboratoires d'avoir accès aux techniques d'analyse du transcriptome. Ainsi, le volume de données d'expression collecté chaque jour dans le monde est colossal et des outils locaux de stockage et d'analyse ont dû être développés [9]. Il paraît alors fondamental de pouvoir cumuler et comparer ces données, même lorsqu'elles sont issues de technologies différentes. Cet échange de données demande actuellement un effort important de standardisation et un formalisme rigoureux.

2.4.1. *Formaliser les connaissances associées au transcriptome : l'annotation syntaxique et fonctionnelle*

Avant d'échanger les données d'expression entre les différents laboratoires, il est essentiel de s'assurer que tous décrivent bien les mêmes objets (c'est-à-dire que les gènes identifiés, et les fonctions attribuées à ces gènes, sont identiques). L'annotation syntaxique des séquences a pour but l'identification des gènes et de leurs éléments régulateurs, l'annotation fonctionnelle s'attache à la détermination de leur fonction. Face à la masse de séquences introduite chaque jour dans les bases de données, ces processus ont été fortement automatisés.

9. Http://genome-www5.stanford.edu, http://cgap.nci.nih.gov/SAGE.

L'annotation syntaxique est un problème extrêmement complexe, surtout chez les eucaryotes à cause de la présence d'introns [10] dans les séquences codantes et de la multiplicité des systèmes de régulation [LEW 00]. Un intense effort d'homogénéisation et de standardisation a été fourni par les bases de données généralistes (GenBank et embl, voir http://www.ncbi.nlm.nih.gov). De plus, des bases spécialisées (dont les annotations sont expertisées) ont été créées pour fournir un maximum d'informations structurées en évitant les redondances. Malgré cela, les descriptions hasardeuses, contradictoires et parfois fausses rendent très difficilement exploitables ces informations par une machine [ILI 03]. De plus, ces données sont en pleine construction et l'arrivée massive d'informations en provenance des centres de séquençage génère une fluctuation très importante des informations (notamment au niveau des identifiants des gènes), ce qui rend pour le moment les échanges et les comparaisons très difficiles.

La principale méthode d'annotation fonctionnelle automatique consiste à rechercher, pour chaque nouvelle séquence, une quelconque homologie avec une autre séquence déjà annotée. Il faut rester conscient des problèmes que cette automatisation peut induire et rester lucide sur la qualité de l'annotation effectuée par cette méthode. En effet, le problème ne réside pas tant dans la comparaison de séquences entre elles, mais plus certainement dans l'interprétation des résultats de cette comparaison. D'une part, les critères permettant de déterminer l'homologue le plus proche sont variés, et, d'autre part, certaines similitudes sont fortuites. De plus, en ce qui concerne les protéines multidomaines, la fonction du domaine présentant le plus haut score d'homologie risque d'être généralisée à la protéine entière, impliquant une confusion voire une omission de certaines fonctions. Enfin, des protéines très similaires peuvent posséder des fonctions très différentes (exemple des paralogues). L'annotation fonctionnelle automatique ne génère donc que des pistes pour l'étude expérimentale.

De même, face à la diversité des relations entre les différentes entités (séquence d'ADN, ARNm et protéines), un effort de description et de formalisation doit être réalisé. Deux protéines peuvent être impliquées dans le même processus cellulaire, que ce soit au niveau d'une même voie métabolique ou dans une même voie de transport ou de régulation, mais peuvent également participer à une interaction spécifique telle qu'une interaction protéine-protéine ou protéine-acide nucléique. Les connaissances décrivant ces interactions sont le plus souvent dispersées dans la littérature scientifique ou dans différentes bases de données très hétérogènes et restent extrêmement difficiles à interpréter de façon automatique.

---

10. Un intron est une séquence non codante dans l'ADN qui vient s'intercaler dans la séquence codante du gène (entre les exons). Cette séquence est éliminée au cours du phénomène d'épissage pour former l'ARN messager. Des épissages complexes peuvent permettre la formation de plusieurs ARNm à partir d'un même gène : il s'agit de l'épissage alternatif.

A première vue, il semblerait que ces problèmes d'annotation des gènes puissent être différés. Il n'en est rien. Lorsque, à l'issue d'une analyse SAGE par exemple, on commence à prendre connaissance des résultats, on se rend très vite compte qu'il est nécessaire de regrouper les gènes selon des critères structuraux et fonctionnels. Ce travail ne peut être réalisé manuellement que pour un petit nombre de gènes, et beaucoup d'informations latentes restent actuellement inexploitées, faute d'outils d'annotation automatique appropriés.

2.4.2. *Formaliser et échanger les données d'expression*

Il est maintenant très clair que les nouvelles technologies ont changé notre vision de l'expérimentation et que les chercheurs doivent être capables de gérer et d'interpréter de grandes masses de données multidisciplinaires et hétérogènes. Cette intégration passe par le développement de modèles de données partagés basés sur des vocabulaires contrôlés, organisés en base de connaissance (ontologies [11]).

Le consortium *Gene ontology* qui s'est mis en place en 1998 propose de produire un vocabulaire contrôlé, applicable à tous les organismes, même si les connaissances associées aux gènes peuvent évoluer et si le rôle des protéines change (http://www.geneontology.org). Il existe ainsi un vocabulaire commun entre des bases de données très différentes (FlyBase (*Drosophila*), *Saccharomyces genome database* (SGD), *Mouse genome database* (MGD) et d'autres spécifiques pour les bactéries ou les plantes), ce qui contribue au développement plus aisé d'outils d'intégration [CAM 03].

Une démarche similaire a été entreprise pour faire face à l'hétérogénéité des techniques permettant l'obtention de données d'expression. A cette fin, un groupe international constitué de biologistes et d'informaticiens a été créé en novembre 1999 (MGED [12]). Un format d'échange international (MIAME [13]) a ainsi été créé et sert actuellement de référence dans la communauté scientifique utilisant la technologie des puces à ADN. Des implantations XML de ce format (MAML et MAGE-ML) ont été développées. Néanmoins, aucune véritable ontologie permettant de décrire des expériences d'analyse du transcriptome n'est actuellement disponible.

La comparaison des données SAGE entre laboratoires pose moins de problèmes. Les résultats sont acquis définitivement, sous forme de fichiers enregistrant les séquences des tags et leur nombre d'occurrences. Toutes les données publiées peuvent être compilées, permettant de comparer le niveau d'expression d'un même tag dans

11. Http://smi-web.stanford.edu/projects/bio-ontology.
12. *Microarray gene expression data*, voir http://www.mged.org/index.html.
13. *Minimal information about microarray experiment*.

tous les échantillons étudiés. La collection la plus complète est installée sur la plateforme *SAGE Genie* (http://cgap.nci.nih.gov/SAGE). Dotée d'une interface *Web* intuitive et de divers outils de représentation, elle réunit plus de cinq millions de tags séquencés dans plus d'une centaine de types cellulaires [BOO 02]. Un exemple d'utilisation simultanée d'une grande masse de données SAGE est la représentation des niveaux d'expression cumulés sur l'ensemble des chromosomes humains [VER 03].

2.5. Conclusion

A la lecture de ce chapitre, le néophyte pourrait avoir une vision très pessimiste de l'analyse du transcriptome avec cette énumération de contraintes, de biais et de limitations dans la description des objets manipulés ainsi que dans les techniques mêmes d'acquisition des données d'expression. Il faut néanmoins conserver à l'esprit que ces techniques, en plein développement, constituent une véritable révolution dans le domaine de la biologie moléculaire ; le nombre croissant de travaux montrant une réelle pertinence dans l'analyse globale des génomes laisse envisager le potentiel diagnostic extraordinaire de ces technologies. De plus, les techniques SAGE et puces à ADN sont devenues, en moins de dix ans, accessibles en routine à la plupart des laboratoires de recherches. Enfin, des collections de données d'expression associées à des données de protéomiques et à des cartes métaboliques ont été intégrées dans des bases de connaissances, permettant ainsi une interprétation des résultats à un niveau physiologique [KAN 02].

Le choix d'une technique d'analyse par rapport à une autre a une influence directe sur la nature des données collectées. La technique SAGE permet une quantification absolue des niveaux d'expression (permettant un partage des données plus facile) et le calcul d'un rapport d'expression entre deux gènes dans deux librairies SAGE différentes ne pose pas de problème de calibrage (si l'on considère que les valeurs proviennent d'un simple dénombrement à l'issue d'un tirage au hasard). La technique des puces ne permet d'accéder qu'à une valeur relative d'expression. Cette limitation de la technique des puces à ADN est également un avantage car la mesure relative du niveau d'expression des gènes est souvent une valeur pertinente par rapport aux questions biologiques posées. Le nombre de gènes du tableau de données est inconnu avant l'analyse avec la technique SAGE ; il est de plus variable et tributaire de l'évolution de l'annotation des génomes (identification des « tags »). A l'inverse, pour la technique des puces, la liste des gènes est fixée et invariable. Cet inconvénient de la méthode SAGE est aussi un avantage puisqu'il permet de travailler sur des génomes partiellement connus afin de découvrir éventuellement de nouveaux gènes impliqués dans le processus étudié.

Peu de travaux ont été publiés proposant une comparaison directe entre les résultats d'analyses par SAGE et puces à ADN. Les comparaisons effectuées simultanément au sein d'un même laboratoire montrent une corrélation satisfaisante [ISH 00],

mais des travaux réalisés indépendamment par des laboratoires séparés peuvent difficilement être exploités. La raison est tout simplement liée au fait que la comparaison reste quasi impossible actuellement par manque de standardisation des techniques et de formalisation des résultats. La standardisation des techniques passe bien sûr par une standardisation des protocoles expérimentaux (standard MIAME par exemple), mais surtout par une planification expérimentale appropriée (définition de témoins positifs et négatifs pour chaque étape par exemple). La formalisation des résultats passe par l'établissement d'ontologies dédiées à la description des données d'expression et au développement d'outils d'analyse et de représentation. Ces derniers aspects sont certainement les verrous les plus cruciaux qui devront être levés dans les années à venir.

2.6. Bibliographie

[BEL 97] BELLIS M., CASELLAS P., « La puce à ADN : un multiréacteur de paillasse », *Médecine/Sciences*, vol. 13, p. 1317-1324, 1997.

[BLA 03] BLAKE W.J., KAERNE M., CANTOR C.R., COLLINS J.J., « Noise in eukaryotic gene expression », *Nature*, vol. 422, p. 633-637, 2003.

[BOO 02] BOON K., OSORIO E.C., GREENHUT S.F., SCHAEFER C.F., SHOEMAKER J., POLYAK K., MORIN P.J., BUETOW K.H., STRAUSBERG R.L., DE SOUZA S.J., RIGGINS G.J., « An anatomy of normal and malignant gene expression », *Proc. Natl. Acad. Sci. USA*, vol. 99, p. 11287-11292, 2002.

[BRA 00] BRAZMA A., VILO J., « Gene expression data analysis », vol. 480, p. 17-24, 2000.

[BRE 00] BRENNER S., JOHNSON M., BRIDGHAM J., GOLDA G., LLOYD D.H., JOHNSON D., LUO S., MCCURDY S., FOY M., EWAN M., ROTH R., GEORGE D., ELETR S., ALBRECHT G., VERMAAS E., WILLIAMS S.R., MOON K., BURCHAM T., PALLAS M., DUBRIDGE R.B., KIRCHNER J., FEARON K., MAO J., CORCORAN K., « Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays », *Nat. Biotechnol*, vol. 18, p. 630-634, 2000.

[CAM 03] CAMON E., MAGRANE M., BARRELL D., BINNS D., FLEISCHMANN W., KERSEY P., MULDER N., OINN T., MASLEN J., COX A., APWEILER R., « The gene ontology annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro », *Genome Res.*, vol. 13, p. 662-672, 2003.

[CAO 02] CAO Z., WU H.K., BRUCE A., WOLLENBERG K., PANJWANI N., « Detection of differentially expressed genes in healing mouse corneas, using cDNA microarrays », *Invest. Ophthalmol. Vis. Sci.*, vol. 43, p. 2897-2904, 2002.

[CAS 98] CASE GREEN S., MIR K., PRITCHARD C., SOUTHERN E., « Analysing genetic information with DNA arrays », *Current Opinion in Chemical Biology*, vol. 2, p. 404-410, 1998.

[CHU 02] CHUDIN E., WALKER R., KOSAKA A., WU S.X., RABERT D., CHANG T.K., KREDER D.E., « Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays », *Genome Biol.*, vol. 3, p. 5, 2002.

[GER 02] GERSHON D., « Microarray technology: An array of opportunities », *Nature*, vol. 416, p. 885-891, 2002.

[HEL 02] HELLER M.J., « DNA microarray technology: Devices, systems, and applications », *Annu. Rev. Biomed. Eng.*, vol. 4, p. 129-153, 2002.

[HOE 03] T HOEN P.A., DE KORT F., VAN OMMEN G.J., DEN DUNNEN J.T., « Fluorescent labelling of cRNA for microarray applications », *Nucleic Acids Res.*, vol. 31, p. E20, 2003.

[HUG 01] HUGHES T.R., MAO M., JONES A.R., BURCHARD J., MARTON M.J., SHANNON K.W., LEFKOWITZ S.M., ZIMAN M., SCHELTER J.M., MEYER M.R., KOBAYASHI S., DAVIS C., DAI H., HE Y.D., STEPHANIANTS S.B., CAVET G., WALKER W.L., WEST A., COFFEY E., SHOEMAKER D.D., STOUGHTON R., BLANCHARD A.P., FRIEND S.H., LINSLEY P.S., « Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer », *Nat. Biotechnol.*, vol. 19, p. 342-347, 2001.

[ILI 03] ILIOPOULOS I., TSOKA S., ANDRADE M.A., ENRIGHT A.J., CARROLL M., POULLET P., PROMPONAS V., LIAKOPOULOS T., PALAIOS G., PASQUIER C., HAMODRAKAS S., TAMAMES J., YAGNIK A.T., TRAMONTANO A., DEVOS D., BLASCHKE C., VALENCIA A., BRETT D., MARTIN D., LEROY C., RIGOUTSOS I., SANDER C., OUZOUNIS C.A., « Evaluation of annotation strategies using an entire genome sequence », *Bioinformatics*, vol. 19, p. 717-726, 2003.

[ISC 02] ISCOVE N.N., BARBARA M., GU M., GIBSON M., MODI C., WINEGARDEN N., « Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA », *Nat. Biotechnol.*, vol. 20, p. 940-943, 2002.

[ISH 00] ISHII M., HASHIMOTO S., TSUTSUMI S., WADA Y., MATSUSHIMA K., KODAMA T., ABURATANI H., « Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis », *Genomics*, vol. 68, p. 136-143, 2000.

[KAN 02] KANEHISA M., « The KEGG database », *Novartis Found. Symp.*, vol. 247, p. 91-101, 2002.

[LEM 98] LEMIEUX B., AHARONI A., SCHENA M., « Overview of DNA chip technology », *Mol. Breeding*, vol. 4, p. 277-289, 1998.

[LEW 00] LEWIS S., ASHBURNER M., REESE M.G., « Annotating eukaryote genomes », *Curr. Opin. Struct. Biol.*, vol. 10, p. 349-354, 2000.

[LIA 92] LIANG P., PARDEE A.B., « Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction », *Science*, vol. 257, p. 967-971, 1992.

[MAS 93] MASKOS U., SOUTHERN E., « A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesized on glass support », *Nucleic Acids Research*, vol. 21, p. 4663-4669, 1993.

[REY 04] REYMOND N., CHARLES H., DURET L., CALEVRO F., BESLON G., FAYARD J.M., « ROSO: Optimizing oligonucleotide probes for microarrays », *Bioinformatics*, vol. 20, n° 2, p. 271-273, 2004.

[ROS 96] ROSS J., « Control of messenger RNA stability in higher eukaryotes », *Trends Genet.*, vol. 12, p. 171-175, 1996.

[ROU 02] ROUILLARD J.M., HERBERT C.J., ZUKER M., « OligoArray: Genome-scale oligonucleotide design for microarrays », *Bioinformatics*, vol. 18, p. 486-487, 2002.

[SOU 74] SOUTHERN E.M., « An improved method for transferring nucleotides from electrophoresis strips to thin layers of ion-exchange cellulose », *Anal. Biochem.*, vol. 62, p. 317-318, 1974.

[VEL 95] VELCULESCU V.E., ZHANG L., VOGELSTEIN B., KINZLER K.W., « Serial analysis of gene expression », *Science*, vol. 270, p. 484-487, 1995.

[VER 03] VERSTEEG R., VAN SCHAIK B., VAN BATENBURG M., ROOS M., MONAJEMI R., CARON H., BUSSEMAKER H., VAN KAMPEN A., « The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes », *Genome Res.*, vol. 13, p. 1998-2004, 2003.

[WEL 90] WELSH J., MCCLELLAND M., « Fingerprinting genomes using PCR with arbitrary primers », *Nucleic Acids Res.*, vol. 18, p. 7213-7218, 1990.

[WIN 03] WINSLOW R.L., BOGUSKI M.S., « Genome informatics: Current status and future prospects », *Circ. Res.*, vol. 92, p. 953-961, 2003.