Developpement d'outils bioinformatiques et statistiques pour l'analyse du transcriptome hepatique humain par puces a ADN : application a la reponse inflammatoire aigue systemique.

Gregory Lefebvre

N° d'ordre: 00000

THÈSE

présentée

devant l'Université de Rouen

pour obtenir

le grade de : Docteur de l'Université de Rouen
Mention Biologie moléculaire

par

Grégory LEFEBVRE

Équipe d'accueil : INSERM - Unité 519
École Doctorale : Chimie-Biologie de Haute-Normandie

Titre de la thèse :

Développement d'outils bioinformatiques et statistiques pour l'analyse du transcriptome hépatique humain par puces à ADN : application à la réponse inflammatoire aiguë systémique.

soutenue le 17 novembre 2006 devant la commission d'examen composée de :

M. le Pr François	Tron	Président
M. le Dr Philippe	Dessen	Rapporteurs
M. le Dr Jean-Loup	Risler	
M. le Dr Dominique	Cellier	Éxaminateurs
M. le Dr Jean-Philippe	Salier	

Il n'y a rien dans ce monde qui n'ait un moment décisif
Henri Cartier-Bresson

À Héléna, Léonie, Éloi,
Avec tout mon amour

Remerciements

Mes premiers remerciements iront à mes deux co-directeurs de thèse, MM. les Dr. Dominique Cellier et Dr. Jean-Philippe Salier, qui ont accepté de co-encadrer cette thèse et qui m'ont chacun témoigné leur soutien et leur confiance tout au long de ces années. Qu'ils trouvent ici l'expression de ma sincère et profonde gratitude.
Un remerciement tout particulier à toi Jean-Philippe qui au quotidien durant ces années m'a toujours prété une oreille attentive sans temps compté et qui par ailleurs n'a jamais douté qu'un jour je terminerai ce travail.
Je préssens enfin que nos chemins se croiseront encore souvent du coté de Bièvres.

J'exprime ensuite mes remerciements sincères à M. le Pr. François Tron directeur de l'unité 519 à l'INSERM, de m'avoir accueilli au sein de son unité et de me faire l'honneur d'éxercer les fonctions de président du jury.

Je remercie également M. le Dr. Philippe DESSEN de m'accorder le privilège d'avoir accepté la fonction de rapporteur au sein de mon jury.

Je souhaiterais pareillement remercier M. le Dr. Jean-Loup Risler de me faire aussi l'honneur d'avoir accepté la fonction de rapporteur au sein de mon jury. Je me rappelle en particulier le premier rendez-vous à Paris accompagné de Jean-Philippe et de Cédric où nous posions ensemble les premières briques du projet et partagions quelques lignes de perl .

Je remercie l'ensemble des membres du jury pour leur participation et leurs conseils.

Je suis aussi reconnaissant auprès de Mme le Dr. Maryvonne Daveau. Merci de ton écoute et de tes conseils. Je n'oublierai pas par ailleurs tes déboires informatiques causant *la mort du petit cheval*.

Mes remerciements sincères également aux membres ou ex-membres de l'équipe :
- À Céline pour l'humeur joyeuse et constante qu'elle apporte tous les matins ;
- À Thierry, mon compagnon de bureau et d'exil, mon ami ;
- À p'tit Rom et son G5. Merci ;
- À Martine pour tous les chocolats suisses ;
- Au club des filles, Frédérique, Gaëlle et Karine ;
- À Cédric enfin, co-équipier promoteur à l'humeur versatile mais à l'amitié constante.

Un remerciement à Abel qui autour des nombreux cafés à la Dina ou au Pebble gâchait systématiquement mes lundis matins en m'interrogeant sur l'avancement de mon manuscript.

Je souhaite également remercier Rob Andrews de m'avoir aménagé du temps afin de terminer ce travail.

Je profite enfin de ces quelques lignes pour remercier ma famille.
Mes chers parents d'abord, je vous remercie infiniment pour votre confiance inaltérable. Rien n'eut été possible autrement.

Ma femme Héléna enfin, à qui revient la meilleure part de ce travail. Pour ton encouragement constant malgré mon découragement parfois, je te suis infiniment reconnaissant.

Table des matières

Première partie

Chapitre 1

Foie et inflammation systémique

Sommaire

1.1 À propos du foie

1.1.1 Anatomie descriptive

Le foie est un organe plein situé dans la cavité abdominale. C'est le plus gros des organes humains.

1.1.1.1 Description macroscopique

Il est entouré par une capsule conjonctive (la capsule de Glisson) qui s'invagine dans le parenchyme hépatique permettant de déterminer des lobes (cf. figure 1.1). Pour comprendre l'organisation générale du parenchyme hépatique, il est indispensable de mettre en place d'abord la vascularisation du foie.

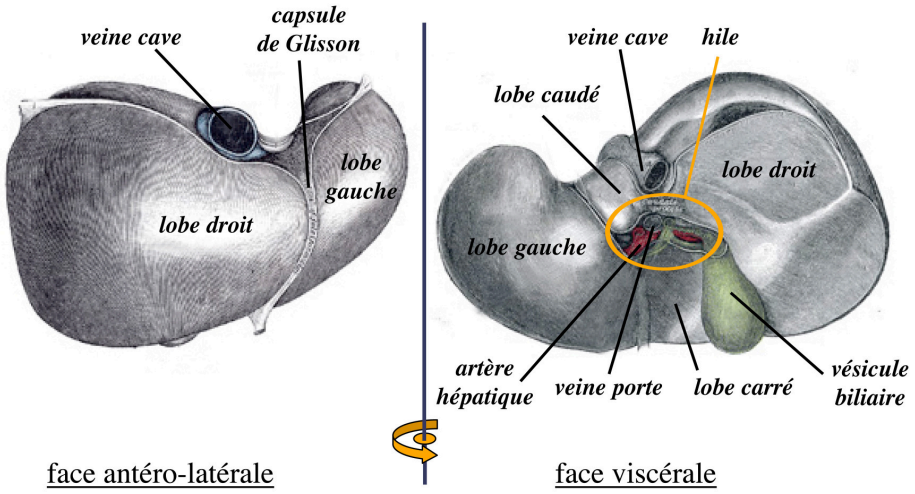


FIG. 1.1 – *Dessin anatomique du foie humain.* La capsule de Glisson partage le foie en quatre lobes : le gauche, le caudé, le droit et enfin le carré. Par la face viscérale les systèmes de vascularisation afférente constituée de la veine porte et de l'artère hépatique, et de vascularisation efférente constituée de la veine cave, pénètrent le foie au sein du hile.

Le foie reçoit deux systèmes vasculaires *afférents* : d'une part la veine porte (70% de la vascularisation) qui draine le sang veineux provenant de la cavité abdominale et particulièrement de la veine mésentérique (intestinale) ; d'autre part l'artère hépatique (30% de la vascularisation). Ces deux voies pénètrent le foie au niveau du hile hépatique et se ramifient pour atteindre les *espaces portes*. Ainsi, les espaces portes ont une signification univoque quant à la nature des vaisseaux qui les composent : ce sont des vaisseaux afférents du foie.

Les veines sus-hépatiques sont les voies *efférentes* du foie et se jettent dans la veine cave inférieure.

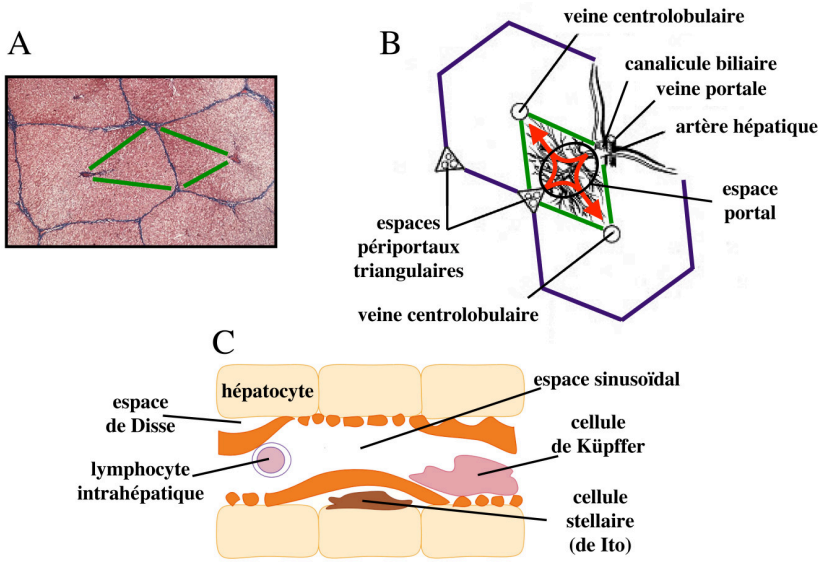


FIG. 1.2 – *Lobules et acinus hépatiques.* Traditionnellement défini comme unité fonctionnelle, le lobule hépatique est surtout très pratique pour décrire l'anatomie du foie. Il est particulièrement mieux dessiné chez le porc (A) que chez l'humain et a une forme hexagonale centrée sur la veine centrolobulaire (B). D'un point de vue physiologique, c'est cependant l'*acinus* qui est réellement l'unité fonctionnelle du foie (losange vert en A et B). Il est centré sur l'espace portal et terminé par les veines centrolobulaires. Les flèches rouges représentent l'orientation de la vascularisation d'abord afférente des espaces périportaux triangulaires puis efférente vers la veine centrolobulaire. Le sang provenant des espaces portes circule le long des sinusoïdes (C). L'architecture de ces sinusoïdes avantage particulièrement les échanges entre les différentes cellules hépatiques et les objets (cellules et molécules) parcourant l'organisme par le système vasculaire.

1.1.1.2 Description microscopique

Du point de vue microscopique, le foie consiste en une myriade d'unités physiquement bien individualisées qui facilitent la description anatomique. Ce sont les *lobules* (cf. figure 1.2-A). Chacun est limité par des espaces portes (irrigués par la veinule porte et l'artériole hépatique) et possède une veinule hépatique terminale centrale (veine centrale du lobule hépatique ou veine centrolobulaire).
Cependant, sur le plan physiologique, la notion d'unité fonctionnelle est attribuée à l'*acinus* (cf. figure 1.2-B) [1]. En son centre se trouve l'espace porte, tandis que les veinules hépatiques terminales occupent la périphérie.
Le sang provenant des espaces portes circule ensuite dans les capillaires sinusoïdes

(cf. figure 1.2-C), limités par les travées d'hépatocytes. Ces capillaires sinusoïdes ont une disposition radiaire et convergent vers la veine centrolobulaire. Leur architecture avantage les échanges entre les cellules hépatiques et les différents objets (cellules et molécules) circulant dans le sang.

1.1.2 Cellules hépatiques

Les cellules de la paroi des capillaires sinusoïdes sont de quatre types [2] : les hépatocytes (cellules endothéliales) ; cellules de Küpffer ; cellules périsinusoïdales stellaires (cellules de Ito) ; et cellules à granulation.

1.1.2.1 Hépatocytes

Cellules polyédriques disposées en travées séparées les unes des autres par les capillaires sinusoïdes, les hépatocytes représentent 80% de la masse du foie et 60% de la population cellulaire. Chaque hépatocyte est baigné par du sang sur deux de ses faces. Leur richesse en organites cytoplasmiques témoigne de leur grande activité métabolique puisqu'en effet ils ont en charge la plupart des fonctions hépatiques.

1.1.2.2 Cellules de Küpffer

Les cellules de Küpffer, fusiformes, sont des macrophages tissulaires. Elles constituent une partie importante du système réticulo-endothélial. Parmi leurs principales fonctions se trouvent la phagocytose de particules étrangères, l'élimination d'endotoxines et d'autres substances nocives, la modulation de la réponse immunitaire par la libération de médiateurs et d'agents cytotoxiques et le présentation de l'antigène [3].

1.1.2.3 Cellules périsinusoïdales stellaires

Les cellules périsinusoïdales stellaires riches en graisses (cellules d'Ito) sont des réserves de dérivés rétinoïdes tels que la vitamine A [4]. Elles se transforment en fibroblastes en réaction aux lésions hépatiques après activation des cellules de Küpffer et jouent ainsi un rôle important dans la fibrose hépatique [5].

1.1.2.4 Cellules à granulation

Les cellules à granulation, qui représentent les cellules les moins nombreuses de la paroi sinusoïdale, sont des lymphocytes granuleux qui agissent comme cellules tueuses naturelles [6].

1.1.3 Rôles essentiels

1.1.3.1 Métabolismes

Le foie est impliqué dans le métabolisme des glucides. En effet, sensible au glucose mais également à d'autre *stimuli* comme l'insuline, les glucocorticoïdes et le système parasympathique, il est le principal capteur de la glycémie. De plus, grâce à des iso-enzymes spécifiques [7], il régule au besoin les voies de la glucogénèse (et/ou néoglucogénèse) et de la glucogénolyse pour respectivement emmagasiner et redistribuer le glucose.
Parallèlement mais dans une moindre mesure si l'on compare au tissu adipeux, le foie intervient sur le métabolisme lipidique dans l'oxydation des acides gras et les synthèses des lipoprotéines et du cholestérol [8] et notamment grâce aux fonctions biliaires du foie.
Par ailleurs, le foie est également en charge de l'emmagasinage de certaines vitamines et particulièrement des composés rétinoïdes [4].
Enfin, le foie intervient dans le métabolisme de détoxication des composés toxiques [9] notamment grâce au cytochrome P450, et permet leur solubilisation et leur élimination dans les urines.

1.1.3.2 Fonctions immunitaires

Le foie reçoit le sang en provenance des intestins par la veine mésentérique. De ce fait, il est particulièrement exposé aux antigènes, ce qui impose des contraintes immunitaires importantes. Celles-ci sont levées principalement d'une part grâce à une micro-architecture fondée sur les sinusoïdes [10] qui facilite les échanges entre le système immunitaire et les cellules hépatiques, et d'autre part grâce à des mécanismes de contrôle afin de déterminer par exemple si l'antigène rencontré nécessite l'amorce d'une réponse immunitaire ou de la tolérance [11]. Le foie possède en effet la faculté de présenter l'antigène notamment grâce aux cellules de Küpffer [3].
Le foie joue enfin un rôle majeur lors de la réaction inflammatoire dans la mesure où les hépatocytes sont responsables des modulations de la synthèse des protéines de la phase aiguë (*Acute Phase Proteins*, APP) [12] et des protéines intracellulaires liées à la phase aiguë (*Acute-Phase-Related Intracellular Protein*, APRIP) [13]. La régulation de la production des protéines de l'inflammation par les hépatocytes est sous la dépendance singulière des cytokines et autres hormones et facteurs de croissance [14, 15].

1.2 Inflammation systémique

1.2.1 Généralités

L'inflammation était déjà connue des Égyptiens et des Sumériens mais elle fût réellement décrite par Celsus au premier siècle de notre ère en ces termes « *rubor et tumor cum calor et dolore* » (rougeur et gonflement avec chaleur et douleur). R. Virchow complétera cet épigraphe au 19ème siècle par les mots « *et functio laesa* (et pertes des fonctions), syndrome apparaissant suite à certaines complications (cf. figure 1.3).

Bien plus tard, Bone [16] affinera cette description dans sa théorie en proposant

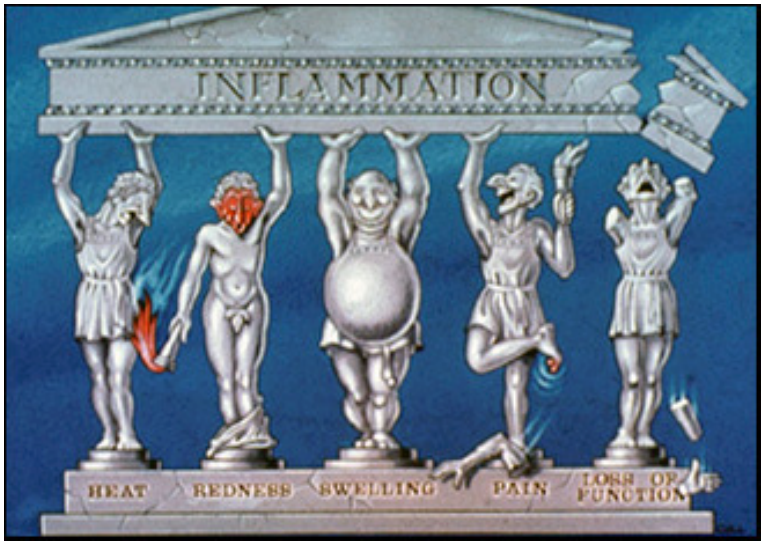


FIG. 1.3 – *Signes cardinaux de l'inflammation.* Ce dessin qui décrit cinq personnages grecs représentant les cinq signes cardinaux de l'inflammation (chaleur, rougeur, gonflement, souffrance et pertes des fonction), s'inspire ainsi des observations de Celsus effectuées 2000 ans auparavant. (P. Cull, *Medical Illustration Department, St Bartholomew's medical college*)

cinq étapes dans l'évolution maligne du syndrome inflammatoire. Ces étapes sont liées à des dissonances croissantes des réseaux des médiateurs de l'inflammation et sont perceptibles cliniquement par des symptômes comme la fièvre ou la tachycardie lorsque les mécanismes de régulation ne parviennent pas à provoquer un retour à *l'homéostasie*.

C'est E. Metchnikoff qui pour la première fois a démontré en 1905 le rôle majeur de l'inflammation dans la résistance aux agents pathogènes [17]. La réaction inflammatoire est une composante de l'immunité innée et c'est la première forme de défense de l'organisme qui doit répondre à une perturbation de l'homéostasie tissulaire [18].

Cette réponse à une agression de l'organisme est stéréotypée et omniprésente, c'est à dire qu'elle est décrite aussi bien chez les mammifères que chez les métazoaires les plus simples. Chez les insectes par exemple, les réactions de mélanisation et l'afflux des hémocytes sur le lieu de la perturbation homéostasique, peuvent être considérés comme des éléments d'une réaction primitive qui permet alors de s'opposer efficacement à de nombreux agents pathogènes [19].
La réaction inflammatoire se décompose classiquement en trois réponses ordonnées et complexes :

- une *réponse locale* initiatrice d'une cascade d'effecteurs primaires et locaux dont les objectifs sont l'élimination rapide de l'agent perturbant ainsi qu'un retour aussi rapide à l'homéostasie ;
- si les effecteurs locaux peinent à apporter le retour à l'homéostasie, l'inflammation entre alors dans une dimension systémique et amplificatrice, c'est la *réponse de la phase aiguë* (*Acute Phase Response*-APR) [12, 20]. Une cascade d'effecteurs est ainsi amorcée et, empruntant la circulation systémique, ces effecteurs orchestrent les réactions des organes protagonistes de cette réponse par l'intermédiaire de leurs récepteurs membranaires spécifiques. On observe alors d'une part un ensemble de modifications physiologiques, biochimiques et nutritionnelles et d'autre part des modifications des concentrations de certaines protéines plasmatiques dont la cause principale est une altération de la transcription des gènes correspondants au sein de l'hépatocyte ;
- la dernière réponse est la *résolution* qui permet un retour à l'homéostasie tissulaire.

Cependant, il arrive que la réponse de la phase aiguë soit déséquilibrée et qu'en conséquence la résolution vers un état homéostasique ne soit jamais atteint. Dans de telles circonstances, la réaction inflammatoire peut s'invétérer en syndrome de réponse inflammatoire systémique (SIRS) accompagné d'un syndrome de dysfonctionnement multi-viscérale (SDMV) [16, 21] et ainsi semer le doute quant au pronostic vital du patient.

1.2.2 Réponse locale

Suite à une rupture de l'homéostasie tissulaire (cf. figure 1.4), les macrophage tissulaires situés à proximité de la lésion sont stimulés par diverses substances comme par exemple des lipopolysaccharides (LPS) des bactéries présentes ou des particules opsonisées.

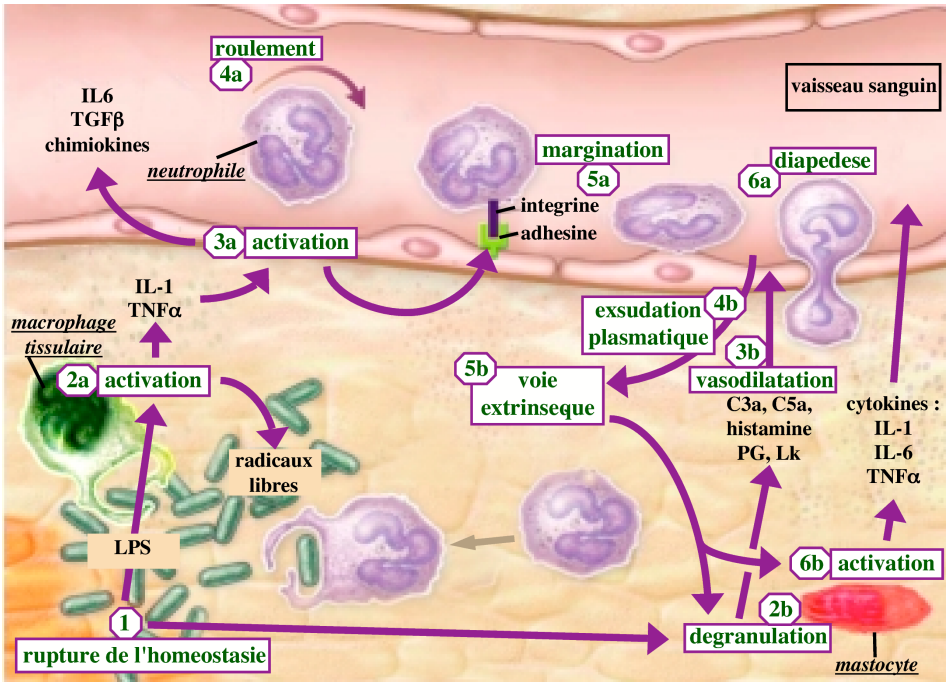


FIG. 1.4 – *Réponse locale de l'inflammation.* Après la rupture homéostasique, deux voies parallèles sont empruntées par l'organisme pour répondre localement à l'agression. La première (2a-6a) fait suite à la réponse des macrophages tissulaires activés, véritables chefs d'orchestre de la réponse, qui libèrent dans le foyer inflammatoire des radicaux libres délétères pour les agents pathogènes et des cytokines dont les cibles immédiates sont les cellules endothéliales et les fibroblastes. Ces derniers sont à leur tour activés autorisant ainsi le recrutement, la margination puis la diapédèse des neutrophiles circulants qui migrent alors au coeur du foyer inflammatoire. La seconde voie (2b-6b) est initiée par la dégranulation des mastocytes lesquels sécrètent alors notamment de l'histamine, responsable de la vasodilatation des capillaires sanguins. Celle-ci facilite la diapédèse des neutrophiles et permet l'exsudation des protéines plasmatiques, ce qui initie la voie extrinsèque qui active les mastocytes. Ces derniers libèrent alors les trois principales cytokines de l'inflammation : l'IL-1, le TNFα et l'IL-6.

Les macrophages libèrent alors principalement d'une part des composés oxygénés tels que des radicaux libres et le monoxyde d'azote (NO) qui sont délétères pour les tissus locaux et les agents pathogènes [22], et d'autre part deux médiateurs, l'interleukine-1 (IL-1) et le *Tumor Necrosis Factor* (TNFα) [15].
Les fibroblastes ainsi que les cellules endothéliales, sensibilisées par l'IL-1 et le TNFα réagissent d'une part en sécrétant des molécules chimioattractantes (chimiokines) dont notamment le *Transforming Growth Factor* β (TGFβ) dans la voie systémique [12, 23], et d'autre part en exposant à leur surface membranaire des adhésines (sélèctines) complémentaires d'intégrines exprimées à la surface des neutrophiles. Ces phénomènes permettent respectivement le recrutement des neutrophiles puis leur *margination* à proximité du lieu de l'inflammation. Les neutrophiles pénètrent alors le milieu extravasculaire par *diapédèse* et atteignent le foyer inflammatoire.

Parallèlement, les mastocytes dégranulés par la perturbation homéostasique libèrent notamment de l'histamine qui est responsable de la vasodilatation des capillaires sanguins. Ce phénomène facilite la diapédèse des neutrophiles et entraîne l'exsudation des protéines plasmatiques suite à laquelle la voie extrinsèque est activée. En conséquence de l'activation de cette voie, un réseau de fibrine est créé et une nouvelle vague de dégranulation de mastocytes est initiée. Enfin l'activation de mastocytes permet la libération dans les voies systémiques de trois cytokines pro-inflammatoires [24] essentielles dans la réponse de la phase aiguë de l'inflammation : l'IL-1, le TNFα et l'IL-6.

1.2.3 Phase aiguë

1.2.3.1 Modifications symptomatiques

Les cytokines libérées dans la voie systémique agissent par l'intermédiaire de récepteurs sur les différents organes protagonistes de la réponse de la phase aiguë de l'inflammation. La moelle osseuse par exemple, sous l'action du *granulocyte-macrophage colony stimulating factor* (GM-CSF) induit la différenciation des neutrophiles et leur remise en circulation. Par ailleurs, l'hypothalamus sous l'action de l'IL-1 et du TNFα induit la fièvre nécessaire à la synthèse des protéines de choc thermique, protectrices endogènes contre les radicaux libres. Mais c'est le foie, et notamment par l'intermédiaire des hépatocytes, qui sous l'action principale de l'IL-6, est le chef de file des acteurs de la réponse [25] lorsqu'il biosynthétise les APP [12] ainsi que les APPRIP [13] qui sont nécessaires pour neutraliser les systèmes activés.

1.2.3.2 Transduction du signal dans l'hépatocyte

Ambivalence des cytokines Les cytokines de l'inflammation sont classiquement organisées en deux catégories ; les pro- et les anti-inflammatoires (cf figure 1.5-A) [26, 27, 28, 29]. Malgré la commodité de cette organisation, cette dichotomie n'est pas franche et les cytokines sont parfois ambivalentes. Ainsi, les

A

Pro-inflammatoire		Anti-inflammatoire
type IL-1	type IL-6	
IL-1	IL-6	IL-6
TNF	HSF	IL-1ra
IFNγ	IFNβ	IL-11
IL-3	LIF	IL-4
GM-CSF	OSM	IL-10
IL-12		IL-13
IL-18		IFNα
		TGFβ

B

FIG. 1.5 – *Ambivalence des cytokines.* Il est classique d'organiser les cytokines en deux catégories : les pro- et les anti-inflammatoires (A). De plus, deux familles de cytokines pro-inflammatoires sont définies en fonction de leur récepteur membranaire spécifique : le famille IL-1 et la famille IL-6. Cependant, comme une pipe peut parfois ne pas être une pipe (B) (R. MAGRITTE, *La Trahison des Images*, 1929), une cytokine pro-inflammatoire peut aussi avoir une activité anti-inflammatoire, à l'image de l'IL-6 (A).

cytokines sont comme certains mots. Seules, elles sont médiatrices d'une action particulière, mais placées dans un contexte (concentration des cytokines, nature de la cellule cible, temps) et/ou associées à d'autres cytokines soit de façon synergique, soit de façon antagoniste, elles sont alors médiatrices d'une autre action [30]. C'est ainsi que l'IL-6 représente l'archétype des cytokines inflammatoires. En effet ses activités sont à la fois pro- et anti-inflammatoires [31, 27, 32], comme si une pipe pouvait à la fois être et ne pas être une pipe (cf. figure 1.5-B). Reprenant pourtant ces deux catégories de cytokines, il est possible de diviser les cytokines « pro-inflammatoires » en deux familles (cf. figure 1.5-A) en fonction du récepteur membranaire auquel elles se lient :

- la famille de l'IL-1, comprenant les ligands des récepteurs IL-1R et TNFαR ;
- la famille de l'IL-6, comprenant les ligands du récepteur IL-6R.

Récepteurs cytokiniques Il existe deux familles de récepteurs membranaires sensibles aux cytokines de l'inflammation.

IL-1R et TNFαR Les IL-1R et TNFαR appartiennent tous à la superfamille des immunoglobulines mais ne sont pas homologues.
Il existe deux types de récepteurs pour les cytokines de la famille IL-1. Le premier (type-I, 80 kDa) est à faible affinité. Cependant la fixation de l'IL-1 à l'IL-1R forme un complexe qui associé à une protéine accessoire (IL-1RAcP) rend la liaison hautement affine. Le second récepteur (type-II, 60-68 kDa) est considéré comme un récepteur factice car son domaine cytosolique est trop court pour transduire le signal [33].
Il existe également deux types de récepteurs pour le TNFα : le type I de 55 kDa et le type II de 75 kDa. Leur affinité envers leurs ligands est comparable et leur niveau basal d'expression dans l'hépatocyte est physiologiquement faible mais s'accroît durant l'inflammation hépatique [34]. Enfin, c'est probablement par l'intermédiaire du récepteur de type I que la synthèse des protéines de la phase aiguë est engagée [35].

IL-6R Les IL-6R sont composés de deux sous-unités α de 80 kDa (IL-6Rα) qui lorsqu'ils lient l'IL-6, complexe deux sous-unités β gp130 qui dimérisées, transduisent le signal [36]. Les récepteurs des autres membres de cette famille ont tous en commun au moins une sous-unités β gp130. Ainsi une cytokine peut se lier à différents récepteurs et un récepteur peut lier différentes cytokines. Ceci peut expliquer la pleiotropie et la bivalence des cytokines de la famille IL-6, car en effet, les cellules n'expriment pas toutes les même sous-unités β et en conséquence cela permet des modulations dans la spécificité des actions des cytokines.
Enfin ces récepteurs n'ont pas d'activité enzymatique au niveau de leurs domaines cytoplasmiques.

Voies de transduction Différentes voies de transduction sont empruntées pour que le signal apporté par les cytokines inflammatoires parcoure le cytosol et pénètre le noyau hépatocytaire (cf. figure 1.6).
Ainsi les signaux délivrés par les IL-1 et TNFα amorcent la voie du *Nuclear Factor κ B* (NFκB), ce qui active le facteur de transcription du même nom [33].
Par ailleurs la fixation de l'IL-6 sur son récepteur amorce la voie des *Janus Kinases - Signal transducers and Activators of Transcription* (JAK-STAT), laquelle se conclue par l'activation du facteur de transcription STAT3 [37, 38].

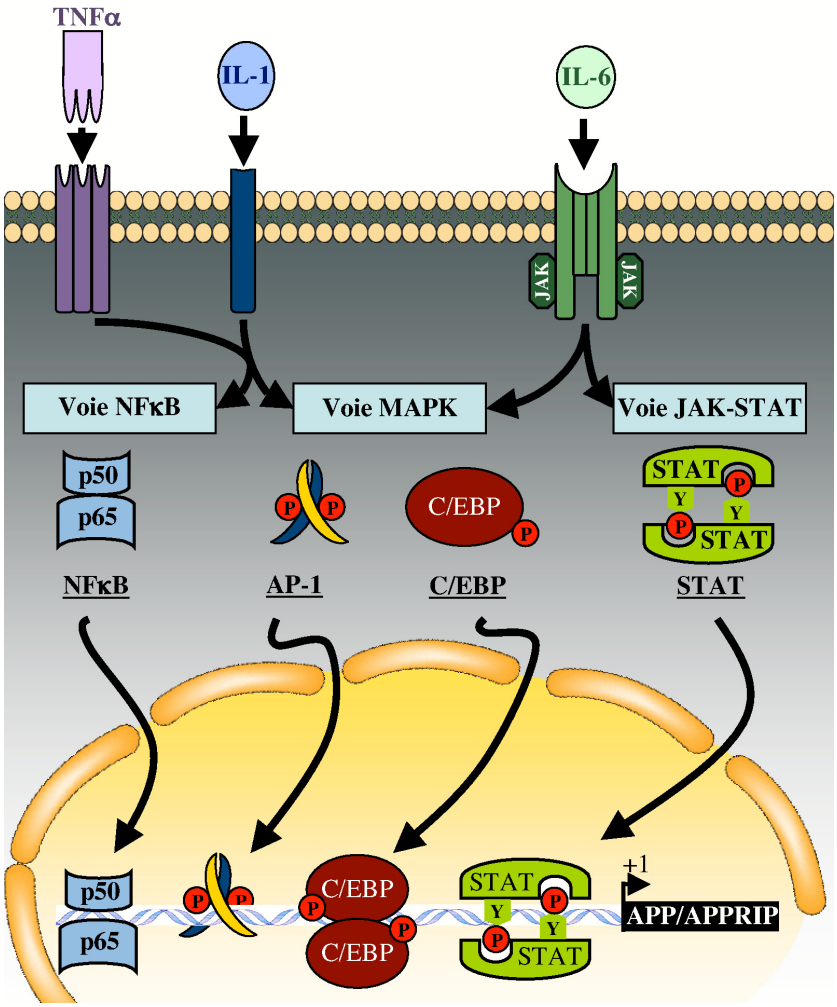


FIG. 1.6 – *Voies de transduction des signaux induits par les cytokines inflammatoires dans l'hépatocyte* La fixation des cytokines sur leurs récepteurs respectifs, localisés dans la membrane hépatocytaire notamment, active différentes voies intracellulaires. Ces voies de transduction du signal des cytokines mènent à l'activation des facteurs de transcription NFκB, AP-1, C/EBP et/ou STAT. Une fois activés, ces facteurs se déplacent dans le noyau cellulaire et se fixent à leurs éléments de réponses localisés au niveau des promoteurs des gènes codant pour les APP et les APRIP.

Cependant ces voies ne sont pas cloisonnées et les signaux des cytokines inflammatoires ont également la possibilité d'emprunter la voie des *Mitogen Activating protein Kinases* (MAPK), laquelle se conclue en activant deux facteurs de transcription : *Activating protein* (AP-1) et *CAAT/Enhancer-Binding Protein* (C/EBP) [28].

Voie NFκB Le NFκB joue un rôle majeur dans le contrôle de l'immunité tant innée qu'adaptative. La famille des facteurs de transcription NFκB est composée de plusieurs membres dont RELA (p65), NFκB1 (p50,p105), NFκB2 (p52,p100), c-REL et RELB. L'architecture de la protéine comprend dans la région N-terminal des domaines de dimérisation, de localisation nucléaire et de fixation à l'ADN. RELA possède en outre dans sa région C-terminale un domaine de transactivation. C'est l'hétérodimère p50-p65 qui est la forme NFκB la plus active [39].
Elles sont présentes dans le cytoplasme et associées à des protéines inhibitrices (IκBα,IκBβ). Les deux événements germinaux de l'activation du NFκB sont la phosphorylation et l'ubiquitylation des IκB qui permettent la dégradation de ces derniers par le protéosome et la levée de l'inhibition portée sur NFκB (cf. figure 1.7). Celui-ci traverse alors la membrane nucléaire et se lie à son site de fixation auprès des gènes dont il régule la transcription [40].

Lorsque l'IL-1 se fixe sur son récepteur de type I, la protéine stabilisatrice IL-1RAcP vient s'associer au complexe. Ce processus permet le recrutement du *MYeloïd Differentiation primary response gene 88* (MYD88) sur une interaction homophilique des domaines *Toll/IL-1-receptor* (TIR) de l'IL-1R et du MYD88 [41], cependant que le domaine N-terminal *Death Domain* (DD) du MYD88 s'associe au *IL-1-receptor-associated kinase* (IRAK) [42]. Subséquent à cette activation, IRAK s'associe au *TNF-receptor-associated factor 6* (TRAF6) nouvellement phosphorylé. Ce dernier complexe alors le *Mitogen-activated protein kinase (MAPK)/Extracellular signal-Regulated protein Kinase Kinase (ERKK)* (MEKK3) [43].

Lorsque le TNFα fixe son récepteur de type I, ce dernier se trimérise et recrute la protéine *TNF Receptor Associated vai Death Domain* (TRADD) qui à son tour s'associe aux protéines *Receptor-Interacting Protein* (RIP), *Fas-as-Associated Death Domain* (FADD) et TRAF2. Ce dernier intervient alors auprès de MEKK3 [44].

MEKK3 est donc une protéine cardinale dans la voie NFκB car elle est d'une part le point de convergence des signaux issus du IL-1R et du TNFαR, et d'autre part une porte vers la voie des MAKP [43]. MEKK3 poursuit la transduction du

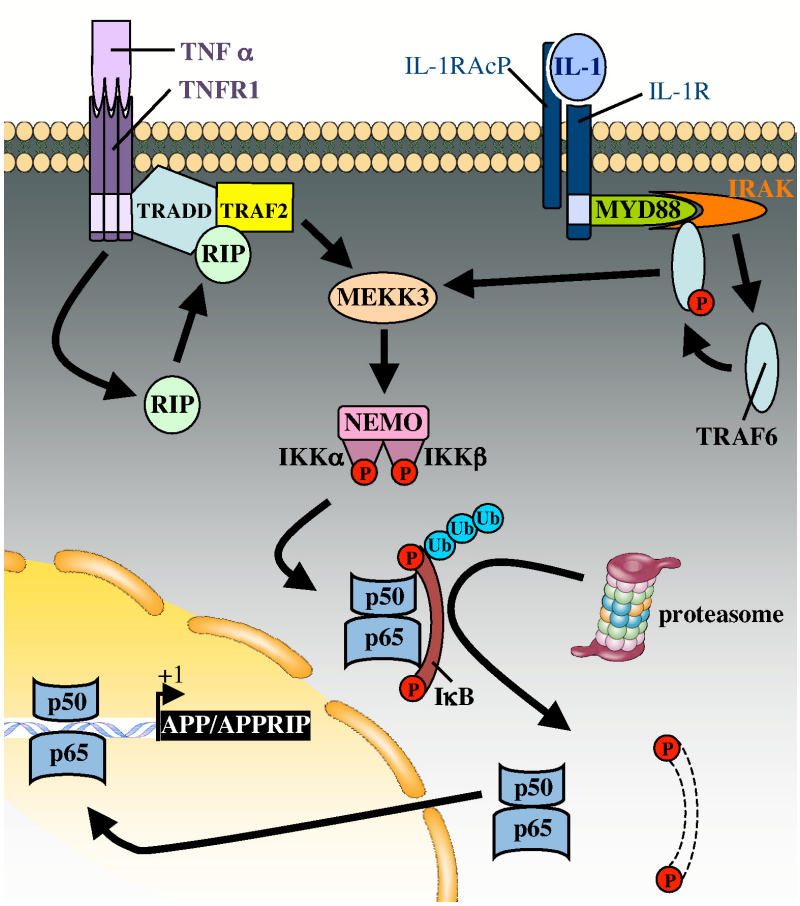


FIG. 1.7 – *Voies du NFκB* Deux voies permettent d'activer la transcription des APP et des APRIP par le NFκB. La première concerne le TNFα qui lorsqu'il fixe son récepteur de type I (TNFR1), ce dernier se trimérise et recrute les protéines TRADD, RIP et TRAF2. Cette dernière intervient à son tour auprès de MEKK3. Parallèlement à la précédente voie, la seconde concerne le signal induit lors de la fixation de l'IL-1 sur son récepteur de type I (IL-1R). Celle-ci provoque la stabilisation du complexe par la protéine IL-1RAcP et permet le recrutement des protéines MYD88 et IRAK. Cette dernière s'associe au TRAF-6 nouvellement phosphorylé qui complexe le MEKK3. MEKK3 activeée, acquière ainsi la capacité de complexe IKK qui phosphoryle IκBα. Ce processus permet la fixation d'ubiquitine [45] et la dégradation de IκBα par le protéasome 26S. NFκB (p65-p50) ainsi libéré, il se déplace dans le noyau afin d'initier la transcription de ses gènes cibles.

signal par la phosphorylation et l'activation du complexe *IκB Kinase* (IKK) qui se compose de deux sous-unités catalytiques (IKKα et IKKβ) et d'une sous-unité régulatrice *NFκB Essentiel MOdulator* (NEMO ou IKKγ) [46]. IKK induit la phosphorylation de IκBα puis son ubiquitylation, ce qui permet sa dégradation par le protéasome 26S. Ainsi libéré de sa protéine inhibitrice, NKκB peut se lier

à l'ADN sur son motif de fixation et activer la transcription de ses gènes cibles (APP et/ou APPRIP).

Voie JAK/STAT La famille des médiateurs JAK est composée de quatre membres : JAK(1-3) et la tyrosine kinase 2 (TYK2). Chacune possède un domaine kinase et un domaine pseudo-kinase régulateur de l'activité kinase de JAK [47]. La famille des facteurs de transcription STAT est composée de sept membres : STAT(1-4, 5A, 5B, 6). La région N-terminale est impliquée dans la régulation de l'activité de STAT et la région C-terminale inclue un domaine *SRC homology 2* (SH2) impliqué dans l'activation et la dimérisation de STAT [48].
Lorsque l'IL-6 fixe son récepteur (cf.figure 1.8), ce dernier se lie de façon covalente à un homodimère de deux sous-unité β gp130 [36]. La dimérisation des β gp130

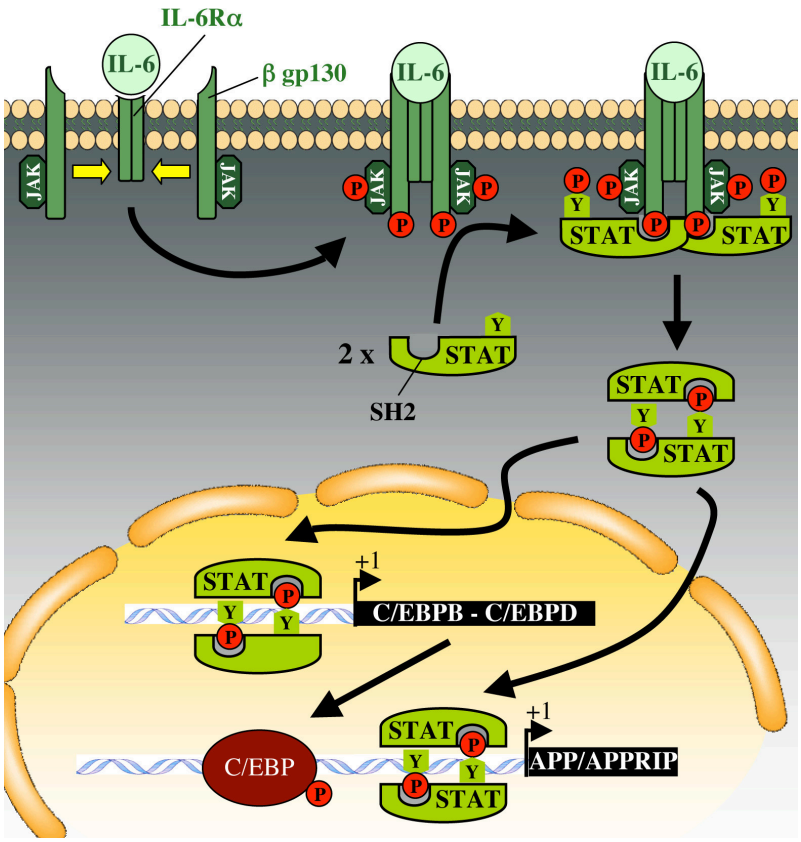


FIG. 1.8 – *Voies de JAK-STAT* La liaison de l'IL-6 à son récepteur induit la création d'une liaison covalente entre celui-ci et un homodimère de deux sous-unités β gp130. Cette dimérisation active les JAK associés au récepteur et recrutent puis phosphorylent les STAT. Ces derniers s'homodimérisent alors grâce au domaine SH2. Ces dimères migrent alors vers le noyau afin d'activer la transcription des APP et APRIP.

est essentielle à l'activation du récepteur mais elle n'est pas suffisante pour initier la cascade de transduction du signal [49] (cf. figure 1.8). Ce sont les JAK qui assurent la phosphorylation des résidus tyrosines (Tyr) des récepteurs et initient ainsi le recrutement des facteurs STAT puis les phosphorylent. Les STAT quittent le récepteur et se complexent en hétéro- ou homodimères grâce à leur domaine SH2 [50, 51]. Ces dimères de STAT migrent vers le noyau et lient l'ADN sur leur motif de fixation afin d'activer la transcription de leur gènes cibles, par exemple des APP et/ou APRIP telles que C/EBPβ et C/EBPδ.

Voie MAPK La voie des MAPK est une cascade de phosphorylations majeure dans la transduction du signal chez les eucaryotes (cf. figure 1.9). En fait, plus de douze MAPK ont déjà été clonés et il est important de comprendre que les cellules possèdent plusieurs voies MAPK, chacune activée préférentiellement en fonction des *stimuli*. Cependant quelque soit la voie MAPK empruntée, on distingue le leitmotiv suivant [52]. Toutes les voies MAPK s'articulent sur trois points :

- Les MAKP sont activées par la phosphorylation concomitante de la Tyr et de la thréonine (Thr). Cette réaction est catalysée par les MEK.
- Les MEK sont activées par la posphorylation des Sérine (Ser) et Thr. Cette réaction est catalysée par les MAP3-kinases (MAP3K).
- Les MAP3K telle que Raf-1 sont activées par des protéines comme celles de la super-famille des Ras et/ou par des oligomérisations et/ou des phosphorylations.

Trois cascades de phosphorylations des MAPK ont été décrites et se concluent sur l'activation des trois MAPK ; *Epidermal growth factor-Regulated Kinase* (ERK), *c-Jun NH_2-terminal Kinase* (JNK) et p38. Les ERK sont activées par les facteurs de croissance et autres mitogènes alors que les JNK et p38 sont induites par des cytokines inflammatoires et le stress cellulaire [53, 54, 55]. La voie des MAPK se conclue sur l'activation des facteurs de transcription AP-1 et C/EBP.

AP-1 est un hétéro-dimère composé des protéines c-FOS et c-JUN. La transactivation des gènes de ces derniers et leur synthèse *de novo* sont contrôlées par les MAPK lorsque celles-ci phosphorylent et activent des facteurs de transcription comme le *Myocyte Enhancer Factor 2C* (MEF2C) et le ElK1.
La famille C/EBP comprend six membres (α-ζ) dont certains sous plusieurs isoformes [56]. Certaines à l'instar de la protéine tronquée LIP ne présentent pas de domaine d'activation mais toutes sont constituée d'un domaine *leucine zipper* de dimérisation et d'un domaine de fixation à l'ADN. L'homo- ou l'hétérodimérisation entre isoforme est nécessaire à la fixation de l'ADN et ce domaine *leucine zipper* est fortement conservé. Au contraire le domaine d'activation est

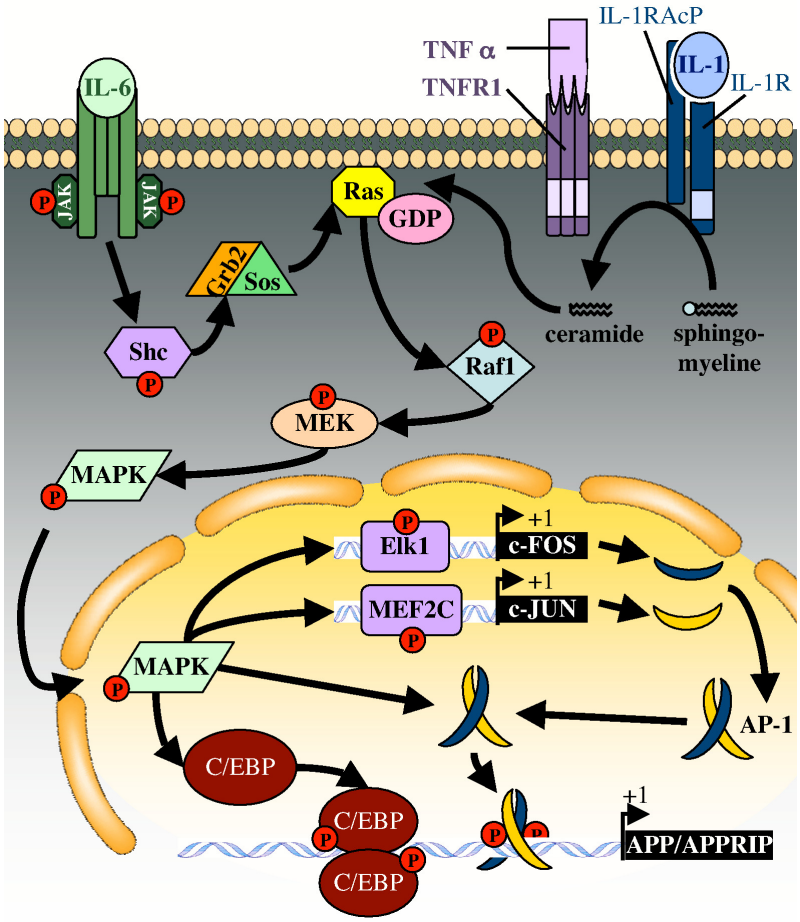


FIG. 1.9 – *Voies des MAPK* La voie des MAPK ets une cascade de phosphorylations dont les MAP3K telle que Raf-1, le MEK et les MAPK sont les protéines essentielles. La ras-GDP est la porte d'entrée principale de cette voie. Elle est activée par les signaux tant en provenance de la famille de l'IL-6R qu'en provenance de la famille de l'IL-1R. Dans le premier cas, la protéine Shc est activée par les JAK du récepteur, ce qui lui permet à son tour d'activer la protéine Grb2. Cette dernière est ainsi capable de s'associer à la protéine Sos qui à son tour active la protéine Ras-GDP. Dans le second cas, c'est la transformation des sphingomyélines membranaires en céramides qui active la Ras-GDP. Le signal emprunte alors la voie des MAPK, lesquelles se déplacent dans le noyau et contrôlent la transactivation ainsi que la synthèse *de novo* de c-JUN et c-FOS par les facteurs Elk1 et MEF2C. Par ailleurs les MAPK activent également les facteurs C/EBP. AP-1 et C/EBP initient alors la transcription des APP ET APRIP.

lui faiblement conservé ce qui est la source d'une régulation de l'action des C/EBP [57, 58]. Dans l'hépatocyte quiescent, la majorité des complexes protéines-ADN contiennent des formes variées d'homodimères c/EBPα et d'hétérodimères C/EBPα-β. Lors de la réponse de la phase aiguë, l'activité transcriptionnelle des

gènes C/EBPB et C/EBPD est augmentée alors que celle du gène C/EBPA est diminuée. Ainsi la participation de C/EBPα dans la dimérisation s'affaiblit au profit des C/EBPβ et C/EBPδ [59].
Dans le cadre de l'inflammation et de l'action des cytokines inflammatoires, l'activation du récepteur de l'IL-6 est capable d'initier la cascade des MAPK par l'intermédiaire des deux protéines associées *Growth factor receptor-bound protein 2* (Grb2) et *Son of sevenless* (Sos) (cf. figure 1.9). Ce complexe formé par l'intervention des protéines *Src homology 2 domain-containing* (Shc), elle-même activée par les JAK du récepteur, active à son tour la protéine Ras-GDP.
Parallèlement, l'activation des récepteurs de l'IL-1 et du TNF entraîne la conversion des sphyngomyélines membranaires en céramides, ce qui active également la protéine Ras-GDP [60].
La voie des MAPK est alors initiée et les MAPK activées transitent vers le noyau. Elles acquièrent la faculté d'activer la transcription de gènes cibles par l'intermédiaire de facteurs de transcription phosphorylés et en conséquence activés. C'est ainsi que c-Fos et c-Jun sont synthétisées et que AP-1 associé à C/EBP initient la transcription de leurs APP et APRIP cibles.

La transduction du signal dans l'hépatocyte emprunte donc trois voies majeures qui forment en réalité un réseau extrêmement complexe et permet ainsi une réponse fine et adaptée de la cellule aux stimuli. Cette complexité vient d'une part du grand nombre de protéines protagonistes qui peuvent tels certains facteurs de transcription agir par synergie ou par antagonisme [61].
D'autre part les trois voies décrites précédemment communiquent entre elles. Ainsi par exemple un grand nombre de facteurs de transcription interagissent avec les C/EBP et notamment C/EBPβ [62, 63].
Enfin la transduction du signal n'est pas linéaire mais suit un modèle séquentiel. En effet l'induction des transcriptions des C/EBPβ et C/EBPδ intervient relativement tardivement après le stimulus inflammatoire. Par ailleurs les transcriptions des facteurs NFκB et STAT3, dont l'activation est plus rapide mais transitoire, sont responsables d'une induction primitive de la réponse puis sont remplacées dans les heures suivantes par les C/EBP [64].

1.2.3.3 Protéines de la phase aiguë

La commission « protéomique clinique » de la Société Française de Biologie Clinique (SFBC) [1] propose cinq critères afin de choisir le meilleur marqueur clinique de l'inflammation :

- une cinétique rapide d'évolution ;
- une augmentation significative du taux plasmatique de la protéine par rapport à son taux basal ;
- une variation indépendante de l'étiologie du syndrome inflammatoire ;
- une dépendance stricte de la réaction inflammatoire ;
- la possibilité d'un dosage précis et rapide.

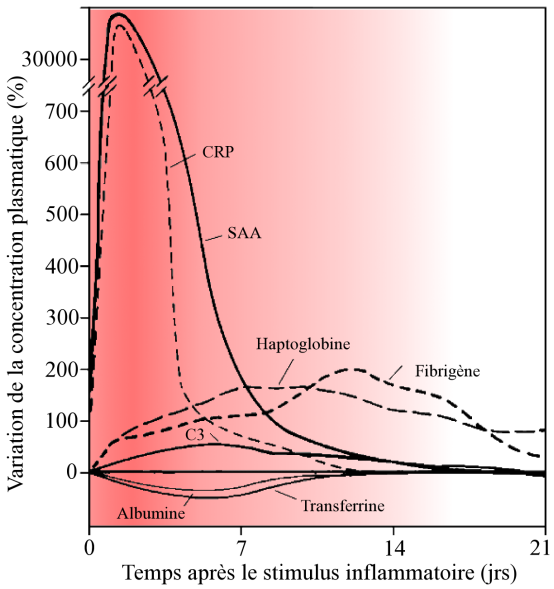


FIG. 1.10 – *Cinétiques caractéristiques des concentrations plasmatiques de quelques APP après le stimulus inflammatoire.* Les deux premiers jours (dégradé rouge croissant) sont marqués par un pic (+30000%) de la concentration plasmatique des APP positives CRP et SAA. Dans un deuxième temps (dégradé rouge décroissant) d'autres APP positives telles que l'haptoglobine, le fibrinogène et le C3 accroissent leur concentration plasmatique (+100%) tandis que des APP négatives telles que l'albumine et la transferrine voient leur concentration plasmatique diminuer (-50%). Enfin vers le $14^{ème}$ jour (fond blanc), l'organisme recouvre une homéostasie.

Cependant selon KUSHNER [65] (cf. figure 1.10), les APP se définissent comme des protéines dont la concentration plasmatique varie d'au moins 25% dans les cinq à sept premiers jours suivant le début du processus inflammatoire aiguë.

[1] http ://www.sfbc.asso.fr/

Cette définition permet une classification bipolaire des APP :
- les APP *positives* dont la concentration plasmatique est augmentée de 2 à 5 fois (C3) voire 100 à 1000 fois pour certaines (CRP, SAA) ;
- les APP *négatives* dont la concentration plasmatique est diminuée de 2 à 5 fois (albumine, transferrine).

Une seconde classification parallèle est également utilisée. Elle permet de classer les APP en deux types selon les cytokines messagères du signal :
- les APP de type I. Leur gène est transcrit après fixation des cytokines de la famille de l'IL-1 sur leur récepteur spécifique ;
- les APP de type II. Leur gène est transcrit après fixation des cytokines de la famille de l'IL-6 sur leur récepteur spécifique.

En effet des régulations croisées interviennent dans les trois voies de transduction précédemment étudiées, et une classification plus ancienne basée uniquement sur la nature sites de fixation de facteurs de transcription en amont des gènes est à proscrire. L'IL-6 par exemple peut dans certaines conditions d'APR, moduler l'activité des C/EBPβ et C/EBPδ. En d'autre termes, l'IL-6 qui normalement initie la voie des JAK-STAT est également capable d'initier la transcription de gènes d'APP I dont le promoteur est dépourvu de site de fixation pour STAT3 [66, 64].

Les APP comprennent d'une part des protéines plasmatiques parmi lesquelles des protéines de transport, des inhibiteurs de protéases et des facteurs de coagulation mais d'autre part des APRIP (cf. tableau 1.1) [13, 67].
La fonction et le rôle des APP au sein du système inflammatoire peuvent être regroupés en trois grands axes :
- épuration et adhérence cellulaire ;
- contrôle enzymatique ;
- coagulation et fibrinolyse.

Épuration L'épuration des composés toxiques ou devenus nuisibles pour l'organisme suite à une lyse cellulaire est un processus qui intervient précocement et de façon importante après le stimulus inflammatoire. Son but est de limiter la détérioration et de préparer la reconstruction tissulaire.
La CRP et le C3 sont par exemple impliqués dans les phénomènes d'opsonisation pour faciliter l'élimination des agents pathogènes par les phagocytes. La CRP sous la dépendance des ions Ca^{2+} est capable de fixer la chromatine, les histones et d'autres ribonucléoprotéines. Elle participe ainsi à la solubilisation des ADN et à leur élimination du foyer inflammatoire. La CRP servirait enfin à éliminer les débris cellulaires et à éviter l'émergence d'antigènes nucléaires à l'origine de maladies auto-immunes [68]. De plus, les endotoxines lipopolysaccha-

APP positives	*Système du complément*	C3 C4
	Système de coagulation et fibrinolytique	Fibrinogène Plasminogène
	Antiprotéases	α-1 protease inhibitor α-1 antichimotrypsine
	Protéines de transport	Haptoglobuline Hémopéxine Céruléoplasmine Orosomucoïde Transferrine *alpha* fetoprotéine
	Autres	C-reactive protein Serum amyloid A α-1 acid glycoprotéine Fibronectine Ferritine
APP négatives	*Protéines de transport*	Albumine Transthyretine Fétuine A (α-2 HS glycoprotéine) Sélénoprotéine P
	Cytokines et facteurs de croissance	Insulin-like growth factor I IL-4
APRIP positives	*Facteurs de transcription*	NFκB IκBα, IκBβ C/EBPβ, C/EBPδ STAT3-5a-b cfos cmyc Glycocorticoïde receptor α
	Protéines cytoplasmiques	Métallothionine 1 et 2 α tubuline β actine cytochrome P450 cRas protooncogène (SH2)-containing protein Y phosphatase
	Protéines membranaires	IL-6R CD14/LPS Binding protein
APRIP négatives	*Facteurs de transcription*	C/EBPα HNF 1α, HNF 3α, HNF 4α cJun
	Protéines membranaires	Growth hormone receptor Glucogen receptor Hepatic LDL receptor

TAB. 1.1 – *Protéines humaines de la phase aiguë.* La grande majorité de ces protéines sont des protéines plasmatiques (APP) néanmoins quelques unes sont des protéines intracellulaires (APRIP). Après un stimulus inlammatoires la concentration de ces protéines dans leur milieu (plasma ou cytoplasme) peut soit augmenter (positives) soit diminuer(négatives) [13, 67]

ridiques bactériennes sont capturées par la SAA [69] qui active de fait le système du complément.

Le foyer inflammatoire est par ailleurs un lieu de séquestration et de lyse des hématies. Il en résulte une libération de molécules d'hémoglobine dans le milieu extra-vasculaire. L'haptoglobine complexe alors aussitôt ces hémoglobines libres et initie en association avec l'hémopexine leur catabolisme au niveau des

hépatocytes et des macrophages. Ce processus contribue au maintien de la sidérémie [70].
Enfin, le stress oxydatif généré par le stimulus inflammatoire est la conséquence d'une libération d'ions superoxydes (O_2^-) dont la neutralisation est sous la dépendance de la céruléoplasmine.

Contrôle enzymatique Une première fonction de contrôle enzymatique consiste en l'inhibition des protéases libérées par la nécrose tissulaire ce qui permet de limiter les dommages subséquents à une activité protéasique effrénée [71]. Ainsi par exemple, l'inhibiteur α1de la trypsine agit sur l'élastase et l'inhibiteur de la chymotrypsine inactive la cathépsine G.
Une seconde fonction de contrôle enzymatique consiste en l'activation du système du complément. Ce système est rendu actif soit par la *voie classique* et réaction antigène-anticorps, soit par la *voie alterne* et des composés issus des micro-organismes. Le complexe CRP-chromatine est également capable d'activer la voie alterne.

Coagulation et fibrinolyse La coagulation sanguine active par l'intermédiaire du facteur XII, la cascade des kinines dont la conséquence est une intensification de la réaction inflammatoire. Les kinines ont une action d'une part chimiotactile auprès des neutrophiles et d'autre part vaso-dilatatrice pour faciliter la diapédèse entre autre.
Au cours de la coagulation, le fibrinogène initie une cascade protéolytique dont le produit est la fibrine. Son accumulation au sein du foyer inflammatoire constitue un lit sur lequel les cellules de l'inflammation peuvent se déplacer. Il est enfin responsable de l'augmentation de la vitesse de sédimentation erythrocytaire, marqueur diagnostique important de la réponse de la phase aiguë.

Les APP négatives La fonction des APP négatives est fortement orientée vers le transport des molécules. Bien que l'utilité de cette diminution de la concentration plasmatique ne soit pas complètement comprise, elle permet au moins d'une part d'atténuer la pression oncotique du plasma due à la sécrétion des APP positives [72] et d'autre part d'augmenter la quantité libre de leur ligand. Or ces ligands (acides gras, hormones, vitamines, oligo-éléments) jouent un rôle important lors de la réponse de la phase aiguë et dans la phase de restauration cellulaire [12].
Enfin la diminution de la concentration plasmatique est une vertu pro-inflammatoires pour certaines APP. Ainsi par exemple la fétuine permet d'augmenter le captage cellulaire d'inhibiteurs cationiques de cytokines pro-inflammatoires [73] et d'initier des mécanismes de désactivation des macrophages [74], sa raréfaction au sein du plasma lui confère une action pro-inflammatoire.

1.2.4 Résolution *ad integrum*

Le retour *ad integrum* circonscrit dans le temps, est la conséquence heureuse de mécanismes de contrôles dont les protagonistes majeurs sont :

- des glucocorticoïdes ;
- des médiateurs lipidiques ;
- des cytokines ;
- des antagonistes naturels et récepteurs solubles ;
- des suppresseurs du signal cytokinique ;
- des facteurs qui influent sur la stabilité du messager.

Glucocorticoïdes La stimulation de l'axe hypothalamo-hypophysaire par l'IL-6 et la production d'ACTH induit une libération de glucocorticoïdes. Ces derniers inhibent la synthèse et la sécrétion de la plupart des cytokines pro-inflammatoires et des cellules stromales [75].

Médiateurs lipidiques Les médiateurs lipidiques anti-inflammatoires peuvent se classer en deux catégories : les lipoxines et les prostaglandines cyclopentonones (cyPG) [29].
Les lipoxines sont générées par une biosynthèse transcellulaire [76]. Elles bloquent le chimiotactisme permettant aux neutrophiles d'approcher le foyer inflammatoire et sont parallèlement pourvues d'un chimiotactisme auprès des mononucléaires libérateurs de cytokines anti-inflammatoires telles que le TGF-β1.
Les cyPG sont les produits de la cyclooxygénase 2 (COX2). Bien que cette enzyme ait une action pro-inflammatoire lorsqu'elle synthétise la prostaglandine E_2 (PGE_2), elle a la faculté de modifier sont processus de biosynthèse afin de produire lors de la phase de résolution des prostaglandines anti-inflammatoires et notamment la 15-désoxy-$\Delta^{12-14}\mathrm{PGJ}_2$ [77]

Antagonistes naturels et récepteurs solubles Les cytokines ont des antagonistes naturels et des récepteurs solubles dont la fonction est de réguler leur biosynthèse. L'antagoniste du récepteur de l'IL-1 (IL-1Ra) fixe l'IL-1 et entre en compétition avec le récepteur de type I mais est dépourvu de mécanisme de transduction du signal [78]. L'induction par l'IL-1 des APP de type I est ainsi inhibée par l'IL-1Ra dans les cellules d'hépatomes, ce qui n'est pas le cas des APP de type II [79, 80].
D'autres récepteurs solubles agissent au contraire comme agonistes à l'image de ceux des IL-6 et CNTF lorsqu'ils recrutent les sous-unités β gp130 [81].

Cytokines Les principales cytokines anti-inflammatoires (cf. figure 1.5-A) ont comme caractéristique fondamentale d'inhiber la biosynthèse des cytokines pro-

inflammatoires comme par exemple l'IL-1, le TNFα.
C'est ainsi que, malgré son activité pro-inflammatoire avérée [31], l'IL-6 entre dans la catégorie des cytokines anti-inflammatoires tant il a été montré son action inhibitrice sur la synthèse des IL-1 et TNFα [82]. L'IL-4 notamment sécrétée par les lymphocytes T *helper* 2 (Th2), régule négativement la synthèse des IL-1, IL-8 et TNFα mais positivement celle de l'IL-1Ra [27].
Enfin, anciennement dénommée « facteur d'inhibition de la synthèse des cytokines », l'IL-10 est la plus importante cytokine anti-inflammatoire connue chez l'humain. Elle possède entre autres les capacités d'une part de désactiver la biosynthèse macrophagique des cytokines pro-inflammatoires, d'autre part d'inhiber l'expression du CD14 qui permet la reconnaissance du LPS et enfin promeut la dégradation des ARMm des cytokines pro-inflammatoires [83].

Suppresseur du signal cytokinique Trois familles de protéines constituent la super-famille des « suppresseurs du signal cytokinique ».
La première regroupe les deux protéines *SH2 containing phosphatases* (SHP-1, SHP-2). La SHP-1 régule négativement le signal de transduction par déphosphorylation. Elle agit grâce au domaine SH2 sur l'IL4-R et les JAK. SH-2 paraît plutôt moduler positivement le signal bien qu'il puisse inhiber sa transduction *via* le récepteur gp130 [84]. La seconde famille est celle des *Protein Inhibitors of Activated STATs* (PIAS). Quatre membres composent actuellement cette famille. PIAS1, PIASx inhibent le facteur STAT1 alors que PIAS3 et PIASy inhibent respectivement les facteurs STAT3 et STAT4. Les deux premiers agissent en empêchant la fixation des facteurs sur leur site alors que l'action des deux autres n'est pas clairement caractérisée [84].
La dernière famille est celle des *Suppressors Of Cytokine Signaling* (SOCS). Sept membres composent cette famille (SOCS 1-7) et toutes possèdent une « boite SOCS » et un domaine SH2. Ce dernier permet une liaison avec les résidus phophotyrosine des JAKs associés aux récepteurs cytokiniques et en conséquence atténue la transduction du signal par compétition entre les SOCS et les STAT [85].

Stabilité du messager Il est désormais bien établi que les changements quantitatifs des APP observés au cours de la phase aiguë ont pour cause majeure des modifications de l'activité péri-transcriptionnelle des gènes correspondant [86]. Cependant depuis longtemps pèse le soupçon d'une régulation également post-transcriptionelle et notamment une régulation au niveau de la stabilité des ARN messagers (ARNm) [87].
La régulation de la stabilité des ARNm est en effet un facteur important dans la modulation de l'expression des gènes. Ainsi l'ARNm de nombreuses cytokines et facteurs de croissance est de courte demi-vie. Cette caractéristique implique la

présence dans leur séquence d'éléments de stabilité dont la nature est très probablement en corrélation avec la fonction biologique de la protéine correspondante [88].
Parmi ces éléments de stabilité, les *Adenylate/urylate (AU)-Rich Elements* (ARE) [89] sont actuellement considérés comme particulièrement intéressants. Ils sont constitués de une ou plusieurs copies du motif *AUUUA* situé en région 3' non traduite (3'-UTR). Ces éléments sont des sites de fixation reconnus par des complexes ribonucléoprotéiques organisés en réseaux (mRNP) telle que la famille de protéines Hu. Ces complexes influencent le devenir et en particulier la dégradation des ARNm [90]. Ainsi l'insertion adéquat dans un gène d'une séquence 3'-UTR riche en ARE, induit une diminution de l'abondance des ARNm du gène [91] réversible par mutagénèse. Enfin la régulation de l'expression des mRNP pourrait être contrôler par des voies de transduction du signal et notamment par la voie p38 MAPK permettant à la cellule un fin réglage du contrôle de la stabilité des ARNm [92].

1.2.5 Transcriptome et inflammation

Au niveau du foie, la réponse de la phase aiguë est un phénomène contrôlé pour l'essentiel par la biosynthèse d'APP et d'APRIP et qui implique de nombreux acteurs. Ceux-ci ont souvent la capacité d'intégrer plusieurs rôles parfois antagonistes en fonction du contexte biologique.
L'ensemble de ces acteurs forme à l'évidence un réseau complexe d'éléments divers et interdépendants. Les réactions de ce réseau aux stimuli inflammatoires se projettent sur deux dimensions :

- une dimension spatiale dans laquelle est mise en valeur la chaîne des éléments qui interagissent pour répondre au stimulus ;
- une dimension temporelle dans laquelle sont mises en valeur les vagues de réponses au stimuli.

Ainsi pour appréhender dans leur globalité les changements de l'expression hépatique des gènes au cours de la réponse de la phase aiguë, il est nécessaire d'élever le niveau de l'étude à l'échelle du génome, *ie* d'étudier le *transcriptome* hépatique. Les *puces à ADN* [93] représentent une famille d'outils qui autorisent cette approche.
Somme toute, peu de travaux sont référencés qui utilisent cette technologie et qui ont pour objet l'inflammation aiguë systémique (cf. tableau 1.2). Qui plus est, trois obstacles apparaissent fréquemment dans la méthodologie.

La premier obstacle concerne le faible nombre de gènes (3500 en moyenne). Ce paramètre est en effet une limite à ce type d'approche dont la vocation est

Référence	**Modèle**	**Spécificité**	**Cardinalité** (gènes)
[96]	Rat	non	7398
[95]	Humain	Foie	4043
[97]	Souris (c57BL/6)	Non	600
[94]	Souris (c57BL/6)	Cytokines	352
[98]	Chien	Non	1000
[99]	Rat	Non	1176
[100]	Rat	Non	8700

TAB. 1.2 – *Puces à ADN pour études de transcriptomes inflammatoires.*

de proposer une vue idéalement exhaustive du transcriptome, or à l'évidence une puce qui correspond à un faible nombre de gènes ne peut remplir cet office.

Le second obstacle qui apparaît dans ces études est celui de la spécificité des gènes représentés. Ainsi seules deux études utilisent des puces dont les gènes sont sélectionnés en fonction de leur tissu d'expression ou de l'activité biologique des protéines correspondantes. La première étude, celle de DONG [94], propose une puce dont la sélection des gènes est orientée vers les cytokines, chimiokines et leurs récepteurs. De même la seconde étude, celle de YANO [95], propose une puce dont la sélection des gènes est orientée vers l'expression hépatique.

Enfin, parmi ces études, seule celle de YANO [95] porte sur l'homme. Cependant, cette étude est fondée sur des échantillons de foie dont les paramètres anatomopathologiques sont mal connus et qui conséquemment induisent des incertitudes dans l'interprétation des résultats. C'est ainsi un point important que de pouvoir déterminer le plus précisément possible les paramètres physiopathologiques des échantillons analysés

L'étude du transcriptome hépatique lors d'une APR peut donc être abordée grâce à la technologie des puces à ADN. D'une part le puce doit être pourvu d'un nombre de gènes suffisamment conséquent pour couvrir l'ensemble du transcriptome. D'autre part, la sélection des gènes posés sur le puce doit être orientée vers les gènes dont l'expression est au moins hépatique. Enfin, la qualité des échantillons et l'accès à leurs paramètres anatomopathologiques sont les fondamentaux d'une analyse fine des résultats. Cependant, en sus des puces à ADN, cette étude doit reposer également sur deux autres socles ; d'une part sur un système d'information permettant d'organiser les données expérimentales et d'autre part des outils d'analyses utilisant des méthodes mathématiques standardisées pour aider à l'interprétation des résultats.

Chapitre 2

Bases de données

Sommaire

2.1 Modélisation orientée objet

2.1.1 Niveaux d'abstraction

L'appréhension d'un problème complexe est facilitée si on le conceptualise et qu'il se présente sous la forme d'un *modèle*. Un modèle est une vue abstraite du monde réel. Il en exprime l'essence grâce à un langage dont la syntaxe est définie et compréhensible. La réalisation d'un modèle passe nécessairement par trois niveaux successifs d'abstraction définis en 1975 par le *American National Standard Institute* (ANSI) [101] mais dont la terminologie peut différer suivant la méthodologie en usage :

- le *niveau physique* ou interne est une représentation abstraite des composants physiques du monde réel ;
- le *niveau d'analyse* ou logique est une représentation intermédiaire entre le niveau précédent et le suivant. C'est au cours de cette étape que seront définies les entités ou *classes* et leur dynamique qui composeront le modèle ;
- Le *niveau de conception* ou externe est la traduction du schéma d'analyse à l'aide du langage de programmation.

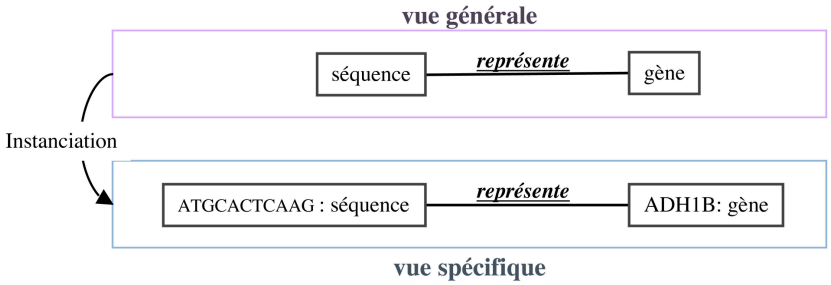


FIG. 2.1 – *Le processus d'instanciation.* La classe *séquence* est associée à la classe *gène* par l'association *représente*. L'instanciation de la classe *séquence* permet de créer l'objet *ATGCACTCAAG*. De même, l'instanciation de la classe *gène* permet de créer l'objet *ADH1B*. L'instanciation est donc le passage d'une vue générale à une vue spécifique du modèle.

2.1.2 Du paradigme objet

La modélisation orientée objet est une approche modulaire de l'algorithmique qui permet de représenter le monde réel sous la forme d'un modèle qui a la particularité d'être constitué d'une collection de concepts uniques [102]. Il est possible de décrire ces concepts selon deux vues, une générale et une spécifique :

- la vue générale les décrit en termes de classes et de relations. Ces dernières forment les abstractions générales ;

– la vue spécifique les décrit en termes d'*objets* et de liens. Un objet est une instance de classe obtenu grâce au processus d'instanciation (cf. figure 2.1). Ces derniers forment les abstractions spécifiques.

Dès qu'il est instancié, un objet entame un cycle de vie durant lequel il bénéficie de connaissances, d'aptitudes et des facultés à communiquer avec d'autres objets par des messages et des stimuli [103]. A la fin de sa vie, l'objet ne sera plus et libérera en conséquence l'espace mémoire informatique qui lui était alloué.
Deux idées fondatrices définissent la modélisation objet. Il s'agit des *caractéristiques* et des *principes primordiaux*. De ces idées fondatrices sont nées des langages de programmation orientée objet et un groupe international pour définir les standards de cette modélisation.

2.1.2.1 Caractéristiques

On définit d'une part les caractéristiques structurelles et d'autre part les caractéristiques comportementales.

Caractéristiques structurelles Les caractéristiques structurelles sont constituées des associations et des attributs.
Les associations déterminent la façon dont les objets communiquent entre eux. Nous verrons plus tard que des notions complémentaires telles la multiplicité et l'orientation peuvent accompagner les associations.
Un objet possède une valeur spécifique pour chacun de ses attributs. Il est à noter que, bien que des objets d'une classe puissent avoir des valeurs égales pour chacun de leurs attributs, chaque objet reste unique et garde sa propre identité. De fait chacun de ces objets bien qu'apparemment identiques, aura une adresse unique dans la mémoire de l'ordinateur.

Caractéristiques comportementales Les caractéristiques comportementales sont constituées des opérations et des méthodes.
Une action qu'un objet peut entreprendre s'appelle une opération et la façon dont il réalise cette opération s'appelle la méthode ou implémentation de l'opération.

2.1.2.2 Principes primordiaux

La dynamique des classes dans un modèle orienté objet est régie par trois principes primordiaux :

– l'encapsulation ;
– la généralisation ou l'héritage ;
– le polymorphisme.

Encapsulation C'est le rassemblement des données et des méthodes au sein d'une même structure ; la classe. En parallèle l'encapsulation masque l'implémentation des opérations à l'utilisateur. Ceci facilite la gestion des modifications et protège les méthodes.

Généralisation C'est la transmission d'attributs et d'opérations d'une classe parent vers une classe fille. C'est une faculté du paradigme objet que de pouvoir utiliser à nouveau des abstractions existantes afin de définir de nouvelles abstractions.

Polymorphisme C'est la possibilité de définir plusieurs méthodes différentes pour une même opération. Deux objets de classes différentes sont donc susceptibles de posséder une opération en commun tout en usant de méthodes différentes. Par exemple, on conçoit aisément que l'opération *validation* d'un objet de la classe *Accepter* n'use pas de la même méthode que celui de la classe *Refuser*.

2.1.2.3 Langages de programmation orientée objet

SIMULA fut le premier langage de programmation implémentant le concept de classes en 1967. D'autres ont suivi tel SMALLTALK en 1976 qui apporta les principes d'encapsulation et de généralisation. Par la suite des langages nativement orientés objets tels JAVA ou C++ sont apparus et ont imposé des langages non-objet dans leur conception à prendre une orientation objet. PERL et PHP par exemple simulent les concepts de la modélisation objet.

2.1.2.4 A propos de l'Objet Managment Group

Dans le but de standardiser la modélisation objet pour l'intégration d'applications hétérogènes, le *Object Management Group* (OMG)[1] fut créé en 1997. Ce consortium international regroupe plus de 850 acteurs du monde informatique tels des constructeurs (IBM, Sun), des producteurs de logiciels (Netscape, Inprise) et des institutionnels (NASA, INRIA). Les concepts généraux mis en avant sont la réutilisabilité, l'intéropérabilité, la portabilité de composants logiciels et l'indépendance face aux systèmes utilisés.
L'élément clef de la vision de l'OMG est le *Common Object Request Broker Architecture* (CORBA)[2] qui autorise la communication de données distribuées

[1] http ://www.omg.org
[2] http ://www.corba.org

indépendamment des types de système d'exploitation, de langage de programmation et des protocoles de réseau. Dans le domaine de la biologie, CORBA est par exemple utilsé par EBI-EMBL afin de faciliter l'acces aux données de leurs serveurs [104]. A noter également le devleppement d'un sysème analogue, le *Distributed Annotation System* (DAS) qui permet à plusieurs groupes de travail repartis sur le globe d'annoter des données centralisées sur un serveur [105].

2.1.3 A propos du Langage de Modélisation Unifiée

2.1.3.1 Historique

Si la méthode MERISE [106] fut la première grande méthode de modélisation développée par les méthodologistes des années 1980, ces derniers se sont confrontés durant cette même période à la naissance d'une nouvelle conception de la programmation informatique ; la programmation orientée objet.
En effet, bien qu'ayant évolué vers la modélisation objet dans sa deuxième version, MERISE a lentement laissé sa place à de nouvelles méthodes nativement objet. Entre 1989 et 1994, une cinquantaine de méthodes de modélisation ont vu le jour, dont notamment les deux méthodes des co-fondateurs de l'*Unified Modeling Language* (UML) [107] : l'*Object-Oriented Software Engineering* (OOSE) de Booch et de Jacobson ; et l'*Object Modeling Techniques* (OMT-2) de Rumbaugh. La description de cette syntaxe me permettra ultérieurement de décrire mes travaux.
En 1994, l'effort autour du langage de modélisation UML est officiellement créé. Certaines entreprises aussi importantes que Hewlett-Packard, IBM, Rational ou encore Texas Instrument comprennent alors l'intérêt de l'UML et créent un consortium. En 1997 d'autres versions sont proposées à l'OMG et en novembre de la même année la version 1.1 des règles syntaxiques de ce langage fut acceptée. Aujourd'hui la version 2.0 d'UML est éditée et maintenue par l'OMG Task Force.

2.1.3.2 Syntaxe

Alors que la sémantique de l'UML est basée sur le paradigme objet (cf. page 36), sa syntaxe basée sur les *diagrammes* [103, 107, 108] est suffisamment souple pour traduire également des modèles relationnels comme les schéma de base de données relationnelles. Un diagramme représente un domaine, c'est-à-dire un ensemble de classes et ses associations. L'UML définit deux grands types de diagrammes :

- les diagrammes structurels ;
- les diagrammes comportementaux.

Diagrammes structurels Les diagrammes structurels facilitent la compréhension de l'agencement des éléments physiques d'un système ainsi que de leur fonctionnalité. Parmi l'ensemble des diagrammes structurels définis proposés par UML, je n'insisterai que sur les *diagrammes de déploiements* et les *diagrammes de classes*.

Diagrammes de déploiement Les diagrammes de déploiement ou diagrammes d'implémentation décrivent l'environnement d'implémentation d'un système et donc le niveau physique de son abstraction. La figure 2.2 illustre un tel diagramme. Ceux-ci possèdent deux types d'éléments :

- les nœuds ;
- les associations de communications.

Les nœuds sont la représentation des ressources disponibles durant l'exécution. Ils sont représentés par un parallélépipède. Un composant résidant dans un nœud est représenté comme emboîté dans ce nœud.
Les associations de communication sont représentées par une ligne continue.

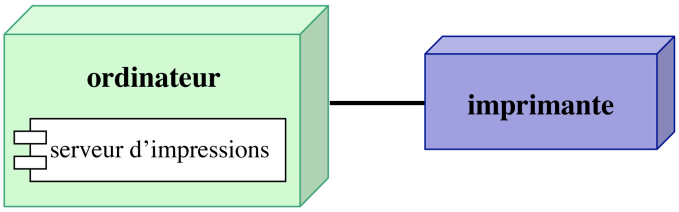


FIG. 2.2 – *Représentation d'une association de communication en UML.* Cette association représente un ordinateur composé d'un serveur d'impression relié à une imprimante.

Diagrammes de classes Les diagrammes de classes dépeignent la structure générale du système et donc le niveau d'analyse de son abstraction. On y retrouve les associations ainsi que les classes, leurs attributs et opérations. Les classes sont représentées en rectangles composés de trois zones (cf. figure 2.3) :

- une zone de titre et de stéréotype de classe, ce dernier étant facultatif ;
- une zone dans laquelle sont placés les attributs ;
- une zone dans laquelle sont placées les opérations. Cette zone ne sera pas représentée dans les classes de stéréotype *table de base de données relationnelle* [109].

Les noms des attributs et des opérations peuvent être précédés par le signe « + » s'ils sont accessibles par les autres classes, le signe « − » s'ils sont aucunement accessibles et par le signe « # » s'ils ne sont accessibles par les classes filles. Par

ailleurs les noms des attributs sont suivis de leur type et les noms des opérations sont suivis du type d'arguments entre parenthèses et du type de retour. Les classes partageant la même utilité peuvent être regroupées sous forme de paquetages qui se représentent par un rectangle surmonté d'un onglet sur la gauche.
De plus UML reconnaît la multiplicité dans l'association, indiquant ainsi combien

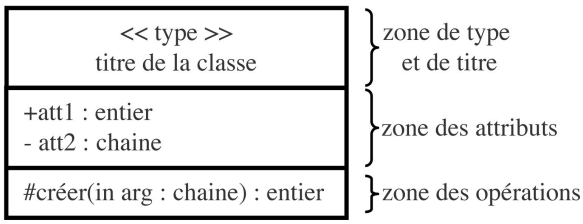


FIG. 2.3 – *Représentation d'une classe en UML.* Dans un diagramme de classe, une classe est représentée par un rectangle lui-même composé de trois zones. La première présente le type et le titre de la classe. La seconde et la troisième présentent respectivement les attributs et les opérations liées à cette classe.

d'objets d'une classe sont potentiellement liés à une autre classe. Les associations peuvent également s'adjoindre un paramètre d'orientation. Enfin l'UML reconnaît quatre types d'associations :

- l'association simple ;
- l'agrégation ;
- la composition ;
- la généralisation.

Une association est binaire si elle associe deux classes ou n-aire si elle associe au moins trois classes. Elle est représentée par une ligne continue entre les classes, à laquelle on adjoint un losange vide central dans les associations n-aires (cf. figure 2.4).

L'agrégation est la caractéristique d'une association qui se traduit par la phrase « possède un ». Elle est représentée par une ligne continue terminée par un losange vide (cf. figure 2.5).

La composition est la caractéristique d'une association qui se traduit par la phrase « se compose de ». Elle est représentée par une ligne continue terminée par un losange plein (cf. figure 2.5).

Une généralisation entre des classes filles et leur classe parent est représentée par une flèche vide continue pointant la classe parent (cf. figure 2.6).

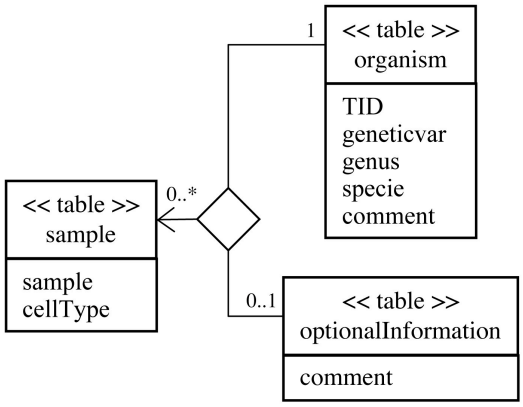


FIG. 2.4 – *Représentation d'une association n-aire en UML.* Les trois classes ont un stéréotype *table* signifiant qu'il s'agit de tables de base de données relationnelle. Les objets *Organism* et *OptionalInformation* sont liées à aucun ou plusieurs objets *Sample*. D'autre part, un objet *Sample* est lié à un seul objet *Enzyme* et à, au plus, un objet *OptionalInformation*.

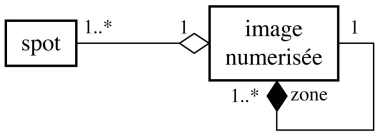


FIG. 2.5 – *Représentation d'une agrégation en UML* Ce diagramme décrit une agrégation ainsi qu'une composition. La première associe le spot comme appartenant à l'image numérisée. La seconde met en lumière le fait qu'une image est composée de zones. Si l'image devait ne plus exister, de fait les zones associées n'existeraient plus.

Diagrammes comportementaux Grâce à la modélisation comportementale, il devient plus facile de comprendre et d'exprimer comment les éléments interagissent. Plusieurs types de diagrammes comportementaux existent, pour autant je ne discuterai ici que des *diagrammes de collaboration*. De plus j'insisterai peu sur ce type de diagramme. En effet dans le cadre d'une modélisation de bases de données relationnelles ces diagrammes sont de moindre importance.

Diagrammes de collaboration Les diagrammes de collaboration expliquent comment les éléments interagissent dans le temps et comment ils sont liés. Ils permettent ainsi de visualiser l'impact d'une interaction dans le temps sur les divers éléments. Les classes, objets et associations sont illustrés à la manière d'un diagramme de classes. La notion de *communication* est apportée grâce à ces diagrammes. Une communication est illustrée sous forme de flèche attachée

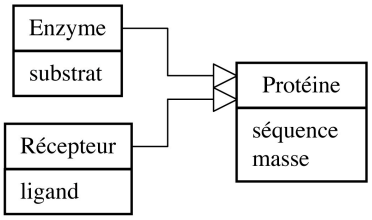


FIG. 2.6 – *Représentation d'une généralité en UML.* Les classes filles *Enzyme* et *Récepteur* héritent des attributs et des opérations de la classe mère *Protéine*.

à la relation, de l'émetteur vers le récepteur. Elle est étiquetée d'un numéro de séquence indiquant l'ordre dans lequel elle est envoyée. La figure 2.7 illustre un diagramme de collaboration.

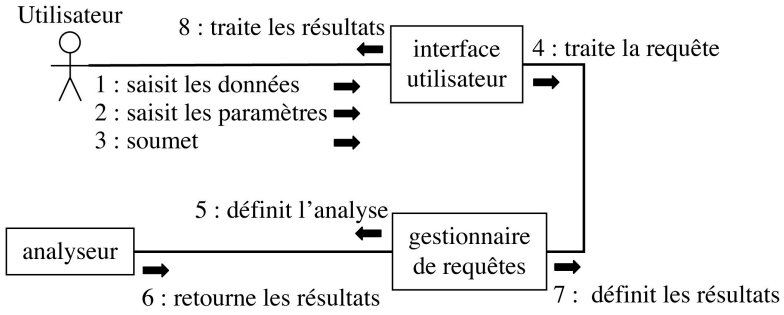


FIG. 2.7 – *Représentation d'un diagramme de collaborations en UML.* L'utilisateur saisit les données puis les paramètres et enfin les soumet. L'interface utilisateur traite la requête, le gestionnaire de requêtes définit l'analyse et enfin l'analyseur retourne les résultats. De façon opposée, le gestionnaire de requêtes définit les résultats, l'interface utilisateur traite les résultats que l'utilisateur peut alors récupérer.

2.1.4 eXtensible Markup Language, un langage très extensible

Alors que le *HyperText Markup Language* (HTML) ne peut capturer la sémantique d'un document électronique et n'est par conséquent pas satisfaisant quant à la transmission de l'information [110], le développement par le W3C[3] du *eXtensible Markup Language* (XML) comble cette lacune.
En effet, version à la fois allégée et évoluée du *Standard Generalized Markup*

[3]http ://www.w3.org

Language (SGML) qui est la référence internationale de description des structures et des contenus de documents électroniques, le XML permet à l'instar de son prédécesseur de consigner la *sémantique* au sein du document grâce d'une part à la création de balises spécifiques et d'autre part à la gestion d'espaces de noms.

Les documents XML ont une structure arborescente qui est définie soit dans une *Document Type Definition* (DTD) ou soit d'une manière plus complexe par le langage schéma XML[4] [111, 112]. Ceux-ci sont fondés sur un modèle objet indépendant du langage de programmation utilisé. C'est ainsi qu'un document XML est constitué d'*éléments* représentant les objets du modèle qu'il suit (cf. figure 2.8). Chaque élément peut contenir un élément fils et des attributs conformément à la terminologie en usage dans la modélisation objet. Chaque élément de l'arbre est en fait une donnée textuelle encadrée par une balise ouvrante et une balise fermante dont la nature révèle la sémantique.

Dans cet environnement, les documents sont d'emblée utilisables par des applications informatiques basées sur l'algorithmique des arbres afin d'en extraire les informations et leur sens. Ainsi le W3C a-t-il développé le *Document Object Model* (DOM) qui est une application recouvrant un ensemble de recommandations pour décrir un modèle objet. Il permet d'accéder aux documents XML et de les utiliser intégralement en tant que structure arborescente.

XML est dores et déjà bien implanté au sein de la communauté scientifique à tel point que nombre de disciplines ont élaboré leur DTD. On référence ainsi :

- le *Mathematical Markup Language* (MATHML)[5] pour les formules mathématiques ;
- le *Chemical Markup Language* (CML)[6] pour organiser les données atomiques, moléculaires, cristallographiques ou encore structurelles ;
- le *Bioinformatics Sequence Markup Language* (BSML)[7] pour organiser les données des séquences biologiques et leurs diverses représentations graphiques ;
- le *Otter Annotation System* qui s'adresse aux séquences biologiques afin de parfaire leurs annotations[113] ;
- le *MicroArray Gene Expression Markup Language* (MAGE-ML) pour représenter les données d'expressions des gènes et que je détaillerai dans la section 2.14 ;
- et d'autres qui sont en développement notamment autour de disciplines telles que la protéomique [108] et la métabolomique [114].

[4] http ://www.w3.org/XML/Schema
[5] http ://www.w3.org/Math/
[6] http ://www.xml-cml.org/
[7] http ://www.bsml.org/

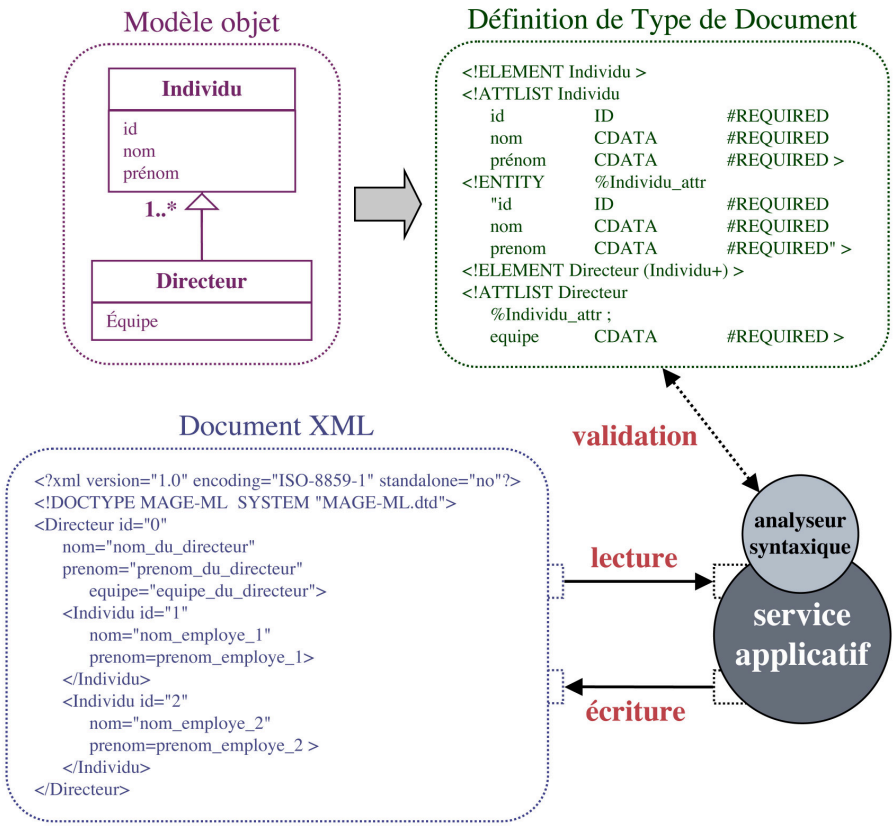


FIG. 2.8 – *Du modèle objet au document.* Traduction d'un modèle objet qu'elle décrit, la *Definition Type Document* (DTD) définit la structure du document XML qui doit être écrit ou lu. Elle permet ainsi à l'analyseur syntaxique associé au service applicatif, de vérifier la structure du document conformément à celle définie dans la DTD. Dans cet exemple, le modèle définit la classe *Directeur* comme héritant de la classe *Individu*. Par ailleurs, un *Directeur* est lié à un ou plusieurs *Individus*. La DTD reprend ce modèle. Chaque classe est représentée par une balise < *!ELEMENT*>. Chaque élément possède une liste d'attributs représentée par une balise < *!ATTLIST*>. La balise < *!ATTLIST*> de l'élément *Directeur* comprend un attribut *%Individu_attr*. Il s'agit d'une entité paramètre qui se réfère à la balise < *!ENTITY*> et simule ainsi l'héritage des attributs de la classe *Individu* par la classe *Directeur*. Le document XML édité est conforme au modèle objet. Le *Directeur* de l'*équipe_du_directeur* est responsable de deux *Individus*.

XML est donc en passe de s'imposer au sein de la communauté bioinformatique et plus généralement au sein de la communauté scientifique tant ses qualités sont intéressantes et ses évolutions prometteuses.

2.2 Gestion des Bases de Données

2.2.1 Architecture à trois strates

Avec l'avènement dans les années 1990 d'Internet et du *réseau des réseaux* (WWW), l'architecture client-serveur a évolué vers une architecture à trois strates (cf. figure 2.9)[115]. La strate *clients* est constituée de multiples utilisateurs qui souhaitent accéder aux données par l'intermédiaire d'un navigateur internet. Elle se charge pour l'essentiel de l'affichage. Cette première strate est liée à la strate *applications* par Internet et le *HyperText Transfer Protocol* (HTTP) couramment utilisé par les navigateurs. Les requëtes de la première strate sont donc réceptionnées par le serveur HTTP de la seconde strate. Le serveur du système applicatif se charge alors d'exécuter le code pour traduire et transmettre les requëtes au système de gestion de base de données (SGBD) de la strate *données*. Ce dernier collecte les informations grâce à un langage de programmation tel le *Structured Query Language* (SQL). Les réponses aux requëtes initiales des utilisateurs parcourent le chemin inverse et sont affichées sur les navigateurs des utilisateurs.

Cette architecture qui présente un déséquilibre prononcé en faveur du serveur

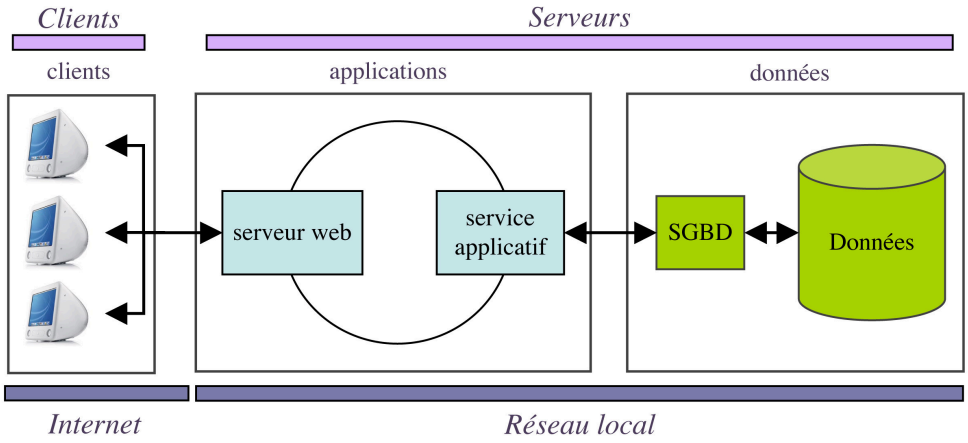


FIG. 2.9 – *Architecture client-serveur en trois strates.* La première est constituée par les clients, la seconde par les serveurs web et d'applications, la troisième par les serveurs de données et les données.

puisque celui-ci accueille à la fois le service applicatif et le SGBD, possède au moins deux avantages. Elle permet premièrement de centraliser l'information et en conséquence de faciliter la maintenance des données au sein de la base. Enfin, l'utilisation d'un serveur HTTP et donc de protocoles et langages standards internationaux, permet à la fois l'accession simultanée à la mëme information par plusieurs clients et une indépendance vis à vis du système d'exploitation utilisé

par les ordinateurs clients.
Un exemple très répandu de ce type d'architecture est celle constituée d'un serveur web Apache[8], du système applicatif PHP[9] et du système de gestion de base de données MySQL[10].

2.2.2 Systèmes de gestion de base de données

À l'image de la programmation informatique, les systèmes de gestion de bases de données ont subi une dichotomie avec d'un coté une optimisation des systèmes relationnels (SGBDR) et de l'autre une évolution des systèmes qui ont alors adopté le paradigme objet (cf page 36).

2.2.2.1 Systèmes relationnels et la normalisation

Systèmes relationnels Le modèle relationnel [116] est basé sur la théorie des ensembles. La manipulation de données est effectuée suivant le concept mathématique de l'algèbre relationnel. Les opérations fondamentales de cet algèbre telles que l'union, l'intersection ou la différence sont la source de nouvelles relations entre les données.
La théorie des ensembles met en œuvre deux notions :

- les domaines ;
- le produit cartésien.

Un domaine est un ensemble fini ou infini de valeurs. On notera ainsi l'ensemble des booléens {0,1} et l'ensembles des acides désoxyribonucléiques {A, C, T, G}.
Le produit cartésien permet la manipulation des données entre différents domaines notés $\mathcal{D}_1, \ldots, \mathcal{D}_i, \ldots, \mathcal{D}_n$. Leur produit noté $\mathcal{D}_1.\cdots.\mathcal{D}_i.\cdots.\mathcal{D}_n$ est l'ensemble des tuples $\mathrm{v}_1.\cdots.\mathrm{v}_i.\cdots.\mathrm{v}_n$, tel que $\mathrm{v}_i \in \mathcal{D}_i$.
La modélisation relationnelle représente les relations à l'aide de tables ou entités dans lesquelles les colonnes sont des attributs repérés par un nom et un domaine, et les lignes sont des tuples. Le nombre de tuples qui composent la relation est appelée la *cardinalité*.
Ce modèle est appelé *modèle à valeurs* car les liens entre les tables ne sont pas des chaînages physiques mais sont des pointeurs logiques [117, 109]. Deux types de pointeurs sont définis. D'une part les clefs primaires qui garantissent l'unicité des tuples et d'autre part les clefs étrangères qui sont les références aux clefs primaires d'une autre table. On désigne par *schéma* d'une base de données relationnelle l'ensemble des relations qui la composent.

[8] http ://www.apache.org
[9] http ://www.php.net
[10] http ://www.mysql.com

pk	Age	Sexe	TubeIdLab	Date	Nom
1	36	M	DUPXXXA	0104191200	DUPONT
2	36	M	DUPXXXB	0211200800	DUPONT
3	72	F	MARXXXA	0009291030	MARTIN
4	72	F	MARXXXB	0312151700	MARTIN

FIG. 2.10 – *La base de données « patient » non normalisée.* Cette table de quatre lignes (tuples) et de six colonnes (attributs) représente la relation *patient*. Deux patients y sont répertoriés et chacun est associé à deux tubes de prélèvements sanguins identifiés par l'attribut *TubeIdLabo*. Cette relation préserve l'intégralité des données et l'intégrité de leurs liaisons. En cela elle répond aux caractéristiques minimales d'une base de données.

La table illustrée par la figure 2.10 préserve l'intégrité des données et permet de recouvrir l'ensemble des données. Par ailleurs, chaque attribut n'appartient qu'à un seul domaine et la clef primaire *pk* de la relation apporte l'unicité de chaque tuple. Elle satisfait donc pleinement la définition d'une base de données. Pour autant cette relation n'est pas sans défauts majeurs. Le premier est la redondance de l'information. Les attributs *Nom*, *Âge* et *Sexe* sont répétés ce qui génère un gaspillage d'espace mémoire.
Le second se porte sur la sémantique. L'attribut *Tube* caractérise la relation et les attributs *Nom*, *Âge* et *Sexe* sont des caractéristiques directe de *Tube*, ce qui est inexact. Cela vient du fait que la relation *patient* ne représente pas un objet unique du monde réel mais bien un mélange de plusieurs objets.
Enfin le dernier concerne la mise à jour des données. Si un patient est entré dans la base sans qu'un tube ne lui soit parallèlement affecté, cela génère une valeur *null* pour la clef primaire *Tube* ce qui est interdit par la contrainte de l'unicité de la clef.
Toutes ces anomalies proviennent du fait que les dépendances fonctionnelles et les formes normales n'ont pas été définies pour cette table (cf. figure 2.11).

Dépendances fonctionnelles Il existe une dépendance fonctionnelle (DF) entre des attributs Att_1 et Att_2, noté $\text{Att}_1 \rightarrow \text{Att}_2$ si connaissant la valeur de Att_1 il ne peut être attribuée qu'une seule valeur à Att_2. Att_1 et Att_2 sont alors respectivement la source et le but de la dépendance. De plus cette dépendance est élémentaire si la source ne comporte pas d'attribut superflu. Elle est enfin directe s'il n'existe aucun attribut C tel que $\text{A} \rightarrow \text{C}$ et $\text{C} \rightarrow \text{B}$.

Normalisation Il existe au moins trois formes normales.

1FN Une relation est sous la première forme normale (1FN) si aucun de ses attributs n'est décomposable.

2FN Une relation est sous la deuxième forme normale (2FN) si elle est 1FN et si toutes les DF entre la clef et les attributs sont élémentaires.

3FN Une relation est sous la troisième forme normale (3FN) si elle est 2FN et si toutes les Df sont directes. La figure 2.11 illustre le devenir de la table patient de la figure 2.10 à l'issu d'une normalisation.

Nom	Age	Sexe	pk_patient
DUPONT	36	M	1
MARTIN	72	F	2

fk_patient	pk_tube	TubeIdLabo	Date
1	1	DUPXXXA	0104191200
1	2	DUPXXXB	0211200800
2	3	MARXXXA	0009291030
2	4	MARXXXB	0312151700

FIG. 2.11 – *La base de données « patient » normalisée.* La relation *patient* de la figure 2.10 est désormais scindée en deux nouvelles relations *patient* et *tube* à l'issu du processus de normalisation.

D'autres formes normales plus avancées existent mais elles augmentent considérablement la complexité du système et imposent l'utilisation de nombreuses tables de corrélations et des jointures subtiles qui engendrent souvent une diminution des performances du système. Le défaut d'impédance du langage SQL est enfin une autre limitation des systèmes relationnels. En effet certains problèmes de cohabitation entre SQL et le SGBDR rendent nécessaires l'adjonction des langages procéduraux tels que C, C++, JAVA, PERL ou enfin PHP.

2.2.2.2 Systèmes orientés objet

Les SGBD qui ont adopté le paradigme objet constituent deux catégories. D'une part on trouve les systèmes intégralement orientés objet et d'autre part les systèmes hybrides relationnels-objets.

Systèmes objets Le premier SGBD objet est une extension du langage objet SMALLTALK (cf. page 38). Le système utilise une structure de données plus complexe que celle utilisée par les SGBDR. Ces structures incluent en effet des poin-

teurs et des tables imbriquées encore appelées collections. De ce fait, le système s'affranchit de la première forme normale.
Le problème majeur de ces systèmes est qu'aucun environnement d'exploitation n'est à ce jour aussi performant que ceux trouvés pour les SGBDR. Par ailleurs les données d'entreprises sont toujours sous la forme relationnelle et aucun principe formel de migration n'a encore été établi.

Systèmes relationnels-objets Enfin il existe les systèmes relationnels-objets qui sont des systèmes hybrides. Apparus en 1992 avec les SGBD UniSQL et Postgres, Informix puis IBM et enfin ORACLE ont accompagné cette évolution. Ces systèmes sont actuellement en rapide évolution animée par deux grands mouvements, d'une part une optimisation interne du moteur du SGBD et d'autre part une promotion des couches réseaux en favorisant les interconnexions d'applications à des SGBD hétérogènes.
Ces SGBD sont donc une extension du modèle relationnel dans le monde de la modélisation objet. Leur structure de données bénéficie des notions de pointeurs, de collections et de méthodes issues du paradigme objet.

2.3 Application aux puces à ADN

2.3.1 Des besoins nourris par l'avancée technologique

Bien que la question de la gestion du recouvrement des données biologiques au sein des bases de données spécialisées soit depuis longue date un sujet d'intérêt [118], l'émergence relativement récente des technologies à haut débit d'analyse telles que les puces à ADN et la génomique fonctionnelle a imposé le développement de bases de données capables d'intégrer des informations dont la taille est à la mesure de celles des génomes dont elles proviennent. Considérant par exemple le projet d'étude désormais classique d'un transcriptome de 10000 gènes sur 30 échantillons dans 5 conditions expérimentales, l'ensemble des données issues de ce projet représentera au moins quelques 2 millions d'unités d'information à emmagasiner. Comment ne pas s'y perdre [119] ?

Pour autant, les espérances et les bénéfices déjà acquis provenant de ces avancées technologiques sont considérables pour la recherche fondamentale et la santé publique. Dans le cadre d'une recherche médicale ou pharmacologique [120], si par exemple la forte expression de certains gènes est corrélée à certains cancers, d'autres conditions affectant cette expression, comme celles tendant à faire diminuer l'expression de ces gènes, peuvent être explorées. Enfin des études comparatives peuvent encore être effectuées comme par exemple la comparaison de profils d'expression de certains échantillons à ceux d'autres échantillons, pour déterminer des ressemblances et dissemblances de profils et ainsi créer un groupage ou classifier des échantillons [121, 122, 123, 124, 125, 126].

Pour déterminer les caractéristiques essentielles nécessaires à l'élaboration de bases de données dédiées aux études d'expression des gènes, il faut de prime abord s'efforcer de concevoir l'ensemble des dimensions portées par la désignation *données d'expression des gènes*. En effet afin de mener à bien l'évaluation et l'interprétation des résultats, il faut se pourvoir nécessairement de cette démarche et c'est ainsi que je distingue quatre dimensions.
D'une part une dimension *technologique* inhérente à la conception des puces à ADN. Ainsi seules des interprétations relatives, *ie* par comparaisons, imposée par l'impossible connaissance de la quantité de matériel biologique déposé sur le puce, pourront être menées.
D'autre part, une dimension *expérimentale* comprenant par exemple les paramètres des conditions d'hybridation, de lavage ou encore de séchage.
De plus, je distingue une dimension *analytique* correspondant aux méthodes d'analyses statistiques dans les étapes de filtrage, normalisation, regroupement et classification.

Je distingue enfin une dimension *informative* relevant des données telles que les annotations géniques et les critères anatomopathologiques. Chacun de ces paramètres influence la valeur des résultats et en conséquence l'interprétation biologique qui en résulte. S'ils ne peuvent être maîtrisés, il convient au moins de les évaluer voire de les connaître.
Il devient dès lors évident que la désignation *données d'expression des gènes* ne peut pas seulement signifier les données de l'intensité des signaux émis, mais elle doit également intégrer la notion de contexte expérimental, *ie* de l'échantillon biologique jusqu'à l'analyse informatisée en passant par les protocoles expérimentaux de la biologie humide. En d'autre termes, dans les études d'expressions de gènes le contexte est *tout* [127]
Il se conçoit donc que le scientifique et le clinicien portent un grand intérêt à posséder un unique outil qui à la fois intègre des données d'expression des gènes et permet leur analyse et les échanges au sein de la communauté scientifique. La conception d'un tel outil impose une réflexion sur l'adoption de certaines caractéristiques pour répondre aux interrogations telles que :

- comment intégrer des données hétérogènes ?
- quels sont les détails des données à enregistrer ?
- comment échanger des données hétérogènes au sein de la communauté scientifique ? Faut-il pour cela définir de nouveaux standards de format de documents, une ontologie ainsi qu'une sémantique ?

2.3.2 Intégration des données hétérogènes

Les données d'annotation des gènes sont hétérogènes car issues de bases de données fondamentalement différentes. Ainsi l'une telle UNIGENE [128] est une base de données fondée sur un algorithme de prédiction de gènes, l'autre telle LocusLink [129] propose des annotations nettoyées et vérifiées, l'autre encore telle SWISSPROT [130] est une base de données nettoyée consacrée aux protéines et l'autre enfin telle KEGG [131] est une base de données renseignant sur les voies métaboliques, les signaux de transduction ainsi que les cycles cellulaires. La question est donc de savoir comment regrouper ces données hétérogènes. Trois méthodes d'intégration des données répondent à cette question :

- les liens hypertext. Ils sont simples à implémenter mais ne peuvent référencer plusieurs sources ;
- l'intégration fédérée. Le schéma de la source doit être intégré dans le schéma global du système. Elle ne nécessite pas de mise à jour. Pour autant deux contraintes sont à prendre en compte. La première est la dépendance aux sites distants qui doit être accessible au moment de l'utilisation de la base. La seconde est la nécessité d'un processus intermédiaire capable de trai-

ter instantanément l'information venant de la source pour la présenter au client ;
– l'intégration matérielle (*datawharehouse*). Les données sont dans ce cas extraites de la source, transformées, épurées, intégrées au système puis associées aux données d'expression des gènes. Pour autant la création d'outils permettant une actualisation automatique des données est indispensable et ainsi affranchie de cette containte, cette méthode est la plus intéressante de par ses capacités d'analyses.

2.3.3 Microarray Gene Expression Data Society

Dès le mois de novembre 1999, la *Microarray Gene Expression Data Society* (MGED) fut créée [127] par les plus importants acteurs et développeurs de la technologie des puces à ADN, tels Affymetrix, l'université de Stanford et le *European Institute of Bioinformatics* (EBI).

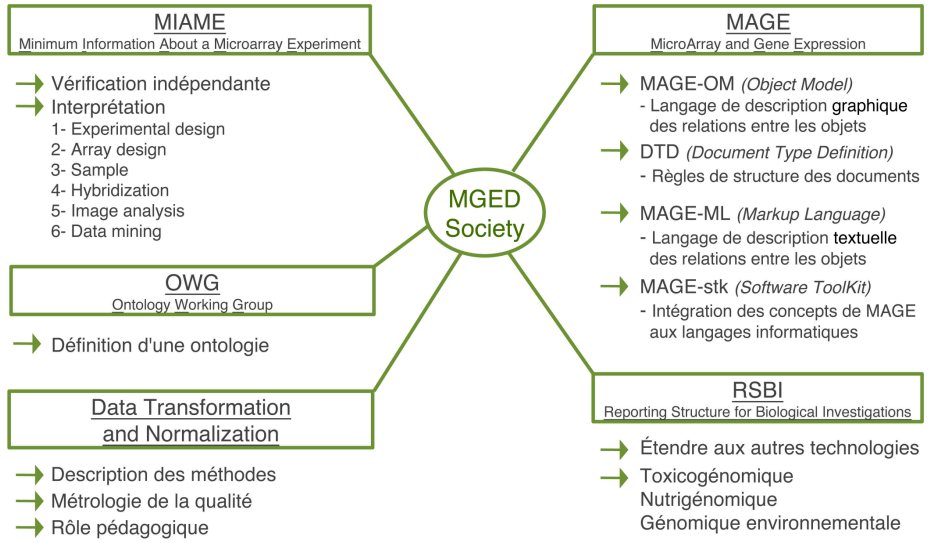


FIG. 2.12 – *La Microarray Gene Expression Data Society.* Cette organisation à but non lucratif dont l'objectif est de créer et de promouvoir des standards autour des données d'expression géniques issues des technologies à haut débit d'analyse, est constituée de cinq groupes : le *Ontology Working Group*, le *Data Transformation and Normalization*, le *Reporting Structure for Biological Investigations*, le *Minimum Information About Microarray Experiment* et le *MicroArray Gene Expression*.

Depuis le mois de juin 2002, la MGED est une organisation à but non lucratif. Son rôle s'inscrit dans la conception et la présentation auprès de la communauté scientifique de standards pour l'annotation et l'échange des données issues des études d'expression des gènes. Ses recommandations facilitent ainsi la création de bases de données et de logiciels implémentant les standards. Bien entendu cette fabuleuse recette fût développée davantage et ses principes adaptés à d'autres disciplines usant de technologies à haut débit d'analyse comme la protéomique [108], la métabolomique [114], la toxicogénomique [132] et ceci de façon non exhaustive [133].

La MGED[11] est constituée de cinq groupes de travail (cf. figure 2.12) dont les trois plus récents sont :

- le groupe *Ontology Working Group* (OWG) qui définit une ontologie spécifique des études d'expression de gènes. Ainsi en fonction de la thématique, des organismes étudiés, des technologies et des méthodes d'analyses utilisées, chacune des données d'expression publiée est rendue univoque grâce à l'utilisation d'un vocabulaire standard ;
- le groupe *Data Transformation and Normalization* qui à la fois élabore une méthodologie pour la description des analyses et une métrologie de la qualité des expériences. Il se charge également de promouvoir ces notions au sein de la communauté ;
- le jeune groupe *Reporting Structure for Biological Investigations* (RSBI) qui prend en charge l'extrapolation de l'ensemble de ces travaux vers les autres technologies à haut débit d'analyse.

Les deux autres groupes mais également les fondateurs de la MGED et en conséquence les plus avancés dans leur travail sont le groupe *Minimum Information About Microarray Experiment* (MIAME) et le groupe *MicroArray Gene Expression* (MAGE).

2.3.3.1 MIAME ou comment bien composer son magasin

Le groupe de travail MIAME [134] est en charge d'émettre des recommandations pour la conception d'un modèle de base de données afin que celle-ci puisse contenir le minimum d'information nécessaire aux bonnes reconduction et interprétation des expériences.
Depuis le mois d'octobre 2002 [135, 136, 137] des journaux tels *Sciences*, *Bioinformatics*, *The Lancet*, *Cell* ainsi que le groupe de publication *Nature Publishing Group* [138, 139] n'acceptent pas d'autres travaux que ceux adoptant les recommandations MIAME. Il est du reste fort probable que d'autres journaux leur

[11] http ://www.mged.org

emboîtent bientôt le pas [140, 141].
MIAME propose un schéma en six paquetages (cf. figure 2.13) concernant :

- les projets expérimentaux (*experimental design*) ;
- les puces à ADN (*array design*) ;
- les échantillons (*samples*) ;
- les hybridations (*hybridizations*) ;
- les mesures (*measurements*) ;
- la normalisation (*normalization*) ;

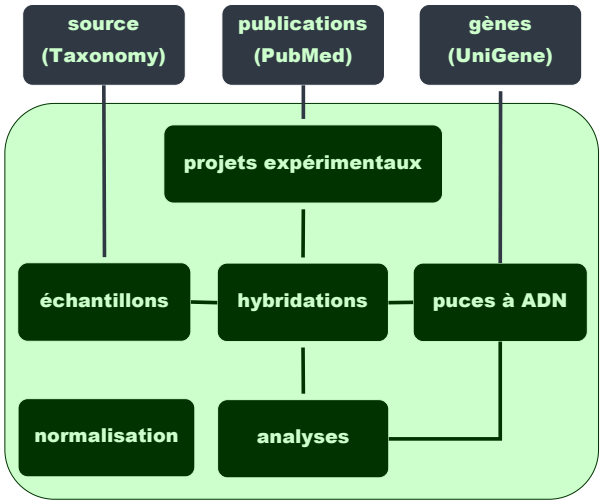


FIG. 2.13 – *Le Minimal Information About Microarray Experiment.* MIAME propose un schéma de base de données constitué de six paquetages dont certains sont associés à des sources externes et concernant les projets expérimentaux, les hybridations, les puces à ADN, les échantillons, les résultats d'analyse et enfin les méthodes de normalisation.

Projets expérimentaux Ce paquetage décrit le projet expérimental dans son intégralité. Il comprend l'ensemble des hybridations menées dans le cadre d'une même question scientifique. Les informations contenues dans ce paquetage concernent notamment les personnes responsables du projet, le type des expériences comme par exemple les cinétiques ou les comparaisons dites *malade versus sain* mais également les facteurs expérimentaux, la liste des organismes et des plates-formes utilisées ainsi qu'une description textuelle du projet. Il décrit enfin les relations entre les échantillons, les puces et les hybridations du projet.

Puces à ADN Ce paquetage décrit d'une part chacune des puces utilisées pour les projets et d'autre part chacune des sondes biologiques déposées sur la puce. Sont donc consignés dans ce paquetage les données telles que le type de puce, le

type de plate-forme ainsi que les caractéristiques, l'identifiant, la nature (produits de PCR, colonies ou oligonucléotides synthétisés *in situ*) et la séquence de chaque sonde de la puce.

Échantillons Ce paquetage décrit les échantillons utilisés pour les projets, la préparation des extraits et protocole de marquage. Il contient donc d'une part des données en terme clinique, de taxinomie et de stades de développement physiologique. D'autre part il contient des informations concernant la préparation des extraits hybridés telles que les amplifications et extractions. Enfin sont également enregistrés dans ce paquetage les protocoles de marquage et les caractéristique du marqueur fluorescent ou radioactif.

Hybridations Ce paquetage décrit les hybridations menées dans le cadre des projets ainsi que leurs paramètres. On y enregistre donc les solutions utilisées, les agents bloquants et autres procédures de rinçage ainsi que par exemple la température et les instruments utilisés.

Analyses Ce paquetage décrit l'ensemble des résultats, de l'image numérisée jusqu'aux données finalisées. Sont donc consignées dans ce paquetage les fichiers bruts issus de la numérisation des puces, les fichiers issus des logiciels de quantification des signaux émis par la puce et enfin les données normalisées et consolidées au besoin par des n-uplets.

Normalisation Ce paquetage décrit les stratégies de normalisation ainsi que les algorithmes utilisés. Il doit également contenir la nature des éléments de contrôle, leur position sur le puce et le protocole d'incorporation dans les extraits hybridés.

2.3.3.2 MAGE ou l'art d'échanger ses données

Recommander un modèle standard de bases de données dédiées aux puces à ADN est certes nécessaire mais ce n'est pas l'alpha et l'oméga. Il convient en effet de proposer également un format standard d'échange des informations. Historiquement, la démarche fut initiée par le développement des *Gene Expression Markup Language* (GEML) [142] et *MicroArray Marckup Language* (MAML) respectivement par la Rosetta Inpharmatics et la MGED. Ces deux langages sont basés sur le XML. En effet ce dernier possède la faculté de transmettre non seulement des données mais également la sémantique qui leur est associée (cf. section 2.1.4). C'est donc dans un effort commun que les développeurs de ces deux langages ont élaboré le *MicroArray Gene Expression* (MAGE) officiellement accepté par l'OMG en 2002 [143].

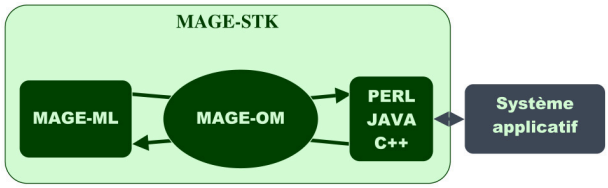


FIG. 2.14 – *Le MicroArray Gene Expression.* Le groupe MAGE propose un modèle objet (MAGE-OM), une DTD (MAGE-ML) et un ensemble d'interface avec les PERL, JAVA et C++ (MAGE-STK).

MAGE est en fait bien plus qu'un simple format de documents puisqu'il regroupe trois composants (cf. figure 2.14) :
- le MAGE-Markup Language (MAGE-ML) ;
- le MAGE-Object Model (MAGE-OM) ;
- le MAGE-Software Tool Kit (MAGE-STK).

MAGE-OM MAGE-OM est le résultat de la modélisation objet des recommandations du MIAME présenté en UML (cf. sous-section 2.1.3). Les paquetages de MAGE-OM reflètent en conséquence les six composants différents du MIAME. En modèle objet complet, il définit des diagrammes structurels ainsi que des diagrammes comportementaux.

MAGE-ML MAGE-ML est un format de fichier semblable au XML, résultant de la traduction du modèle MAGE-OM par respect de règles définies dans la DTD. Chaque classe du MAGE-OM est représentée par un élément dans le document, à l'image d'un document XML classique, avec des attributs et des éléments fils.

MAGE-STK Afin de permettre aux différents laboratoires de répondre aux exigences du MIAME et d'échanger leurs documents MAGE-ML, la suite logicielle MAGE-STK fut développée. Il s'agit d'une interface d'applications (une API) qui permet d'utiliser MAGE au travers de langages de programmations comme PERL, JAVA et C++. Cette interface permet donc la lecture et l'écriture de document MAGE-ML à partir d'un système applicatif quelconque intégré par exemple dans une base de données relationnelle. Cependant il n'existe pas de règles officielles de translation de MAGE-OM vers une base de données relationnelle et il est donc nécessaire d'adapter ces APIs localement.

Bien que MAGE ait été développé dans le cadre de l'élaboration d'un grand conservatoire à l'image de ArrayExpress à l'EBI afin d'accueillir l'ensemble des données de *la communauté des puces à ADN*, ce modèle reste tout à fait applicable à des structures d'échelle plus petite telles qu'un laboratoire tant il est flexible.

2.4 État de l'art

Annuellement, le journal *Nucleic Acid Research* publie au mois de janvier la collection des bases de données publiques jugées utiles au biologiste moléculaire. En 2004, 548 bases étaient référencées et en janvier 2006 la liste en comportait 858 [144, 145]. Des bases de données en nombre toujours croissant et par ailleurs très utilisées [145].
Dans cet ensemble, certaines sont dédiées aux puces à ADN mais elles ne sont pas toutes équivalentes en terme d'importance. Je distinguerais d'une part les bases de données dédiées à une thématique et d'autre part les conservatoires publiques. Enfin, en sus de cette collection, je présenterai également une sélection de *Laboratory Information Managment Systems* (LIMS) qui sont des logiciels dédiés à une utilisation au sein d'un laboratoire indépendamment de la thématique développée autour des puces à ADN.

2.4.1 Bases de données dédiées

Ce sont des bases de données dont le but est de mettre à la disposition de la communauté les données d'expression géniques spécifiques d'une ou plusieurs thématique du laboratoire d'accueil. Elles sont ainsi supposées permettre leur récupération ainsi que leur comparaison à d'autres données issues de diverses sources. Les données contenues sont donc rendues publiques mais le dépôt est souvent impossible. Par exemple la *Stanford Microarray Database* (SMD) qui est la plus importante base de ce type [146], donne accès publiquement aux résultats des travaux menés à l'université de Stanford mais il est impossible d'y déposer des données sans être chercheur ou collaborateur de l'université. De plus, développée sous licence libre, la SMD est téléchargeable gratuitement. Elle est cependant construite sur un SGBD difficile à installer au sein d'une structure de petite ou moyenne importance.

D'autres bases de données sont référencées mais elles sont à la fois plus spécialisées et moins importantes (cf. table 2.1). Il s'agit en effet de bases dédiées spécifiquement par exemple à des organes comme la *Kidney Development Database* pour le rein, la *Brain Gene Expression Database* (BGED) pour le cerveau, ou les HugeIndex et GeneNote pour les tissus sains en général. Il peut encore

Noms	Thématiques	Outils d'analyse	MIAME	Format d'échange
SMD [146]	Données brutes Données normalisées	oui	oui	html text (ftp)
SOURCE [147]	Données vérifiées Sources externes	oui	oui	html text (ftp)
BodyMap	Données vérifiées Tissus humains et murins	non	non	html
CleanEx	Données vérifiées Tissus divers	non	non	html text (ftp)
GeneNote	Données vérifiées Sources externes Tissus humains sains Affymetrix	non	non	html

TAB. 2.1 – Présentations de quelques bases de données dédiées et leurs caractéristiques

s'agir de bases comme la BodyMap qui regroupe les données des niveaux d'expression des gènes orthologues murins et humains. Enfin d'autres à l'image de la SOURCE [147] ainsi que de la CleanEx permettent de rechercher des données d'expression de gènes à travers plusieurs sources après que celles-ci aient été épurées et vérifiées.
L'ensemble de ces bases se donne l'objectif de présenter et de rendre publique leurs données, pour autant aucun moyen n'est vraiment développé pour faciliter la récupération et la ré-analyse des données. Ainsi seules les SMD, SOURCE et CleanEx proposent un serveur FTP pour récupérer l'ensemble des fichiers au format texte. Cependant chacune de ces dernières possède son propre formatage et aucune ne s'accorde afin de proposer un format standard tel MAGE-ML pour exporter les données. Ainsi intégrer automatiquement et localement ces données nécessite une transformation pour laquelle il doit être créé une application adaptée spécifiquement à chaque base dédiée.
C'est ainsi la démonstration de la puissance des recommandations de la MGED que de proposer un conservatoire centralisant les informations dans un format général et standard nécessitant en sorte toujours la même transformation des données dans le modèle MAGE-OM [148].

2.4.2 Conservatoires publiques

Les conservatoires publiques sont des structures capables de recevoir l'ensemble des données d'expression des gènes issues non seulement des puces à ADN mais également d'autres technologies telles le SAGE. Leur objectif est de centraliser les données et de les rendre publique à l'image des DDBJ/EMBL/GenBank pour séquences moléculaires. Ils deviendront sans nul doute une étape incontournable dans le processus de soumissions des travaux scientifiques aux périodiques. Je présente ici les trois grands conservatoires que sont d'une part le *Gene Expres-*

sion Omnibus (GEO), d'autre part le ArrayExpress et enfin le récent *Center for Information Biology Gene EXpression* (CIBEX).

Gene Expression Omnibus Hébergé au National Institute of Health (États-Unis), ce projet fut initié en 2000, avant la lettre ouverte de la MGED, afin de répondre à la demande grandissante d'un conservatoire publique capable d'emmagasiner et de publier l'ensemble des données d'expression des gènes. Avec ses 28000 expériences répertoriées en décembre 2004, GEO [149] est aujourd'hui le plus important conservatoire publique. Après modification, le schéma de GEO est aujourd'hui conforme au schéma MIAME. Il n'est ainsi pas interdit de penser que GEO adoptera bientôt le format d'échange de fichiers MAGE-ML à la place de l'actuel *Simple Omnibus FormaT* (SOFT).

ArrayExpress Le ArrayExpress [150] est hébergé par le EBI (Royaume-Uni) et fut initié deux années après le projet GEO. Il est le premier conservatoire de données d'expression des gènes dont le schéma est basé sur MIAME et le format d'échanges des fichiers sur MAGE-ML. Ainsi deux voies de soumissions des données au format MAGE-ML sont proposées, soit par une interface web nommée MIAMExpress qui dans un un premier temps accepte des données non formatées et dans un deuxième temps les formatent puis les dépose, soit par le dépôt des fichiers pré-formatés grâce au LIMS du laboratoire émetteur.

Center for Information Biology Gene EXpression Encore en cours de développement, le CIBEX [151] est hébergé au National Institute of Genetics (Japon). C'est également un conservatoire publique développé à l'image du ArrayExpress, c'est à dire conforme aux recommandations de la MGED dont le but est d'entreposer les données des études d'expression des gènes mais également celles issues de spectrométrie de masse venant de la zone Asie.

Last but not least, ArrayExpress impose donc son schéma de base de données et son format d'échange de données aux autres conservatoires publiques. Par ailleurs un effort de collaboration entre ces trois conservatoires est du reste sous-tendu par les différents protagonistes. C'est donc avec efficacité que la MGED promeut ses concepts et ne tardera pas, sans nul doute grâce aux périodiques, à convaincre l'ensemble des communautés des technologies à haut débit d'analyse.

2.4.3 Du Laboratory Information Managment System

Un LIMS est un logiciel développé afin d'organiser et d'enregistrer au quotidien les paramètres expérimentaux dans un laboratoire. Qui plus est, il doit également potentiellement aider à la décision en proposant des outils d'analyse.

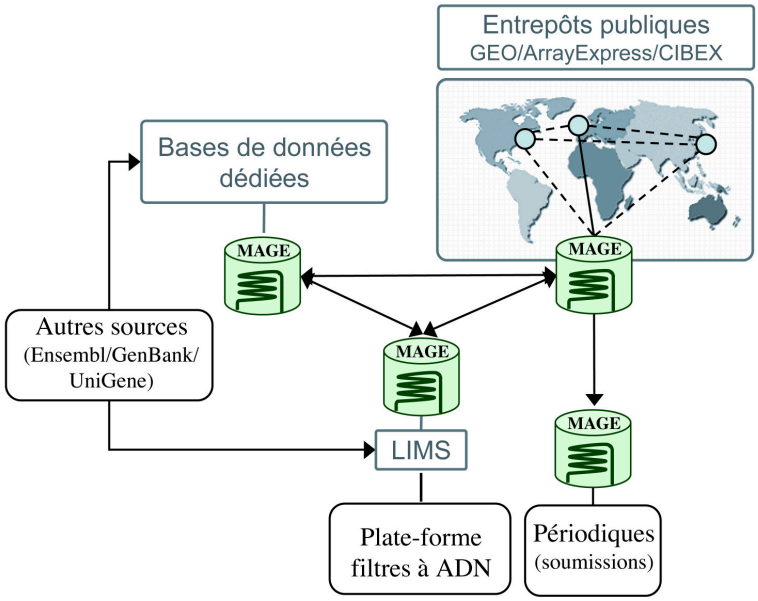


FIG. 2.15 – *Agencement possible entre les différentes bases de données.* L'ensemble des transactions entre les différentes sources de données de puces à ADN est effectuée idéalement par l'intermédiaire du MAGE. Ainsi, associés aux plate-formes de puces à ADN des laboratoires, les LIMS peuvent d'une part récupérer les informations soit auprès de sources diverses telles UniGene soit auprès de bases de données dédiées (idéalement grâce au MAGE), et d'autre part déposer les résultats d'expériences dans les conservatoires publiques. Ces derniers permettent alors aux auteurs de soumettre leurs travaux aux périodiques scientifiques.

Dans le cadre des puces à ADN, il est de plus impératif que le LIMS interagisse avec différentes sources de données et les conservatoires publiques (cf. figure 2.15). Les développeurs de LIMS sont extrêmement prolixes et la table 2.2 présente une liste évidemment non-exhaustive des produits de leur créativité. L'offre ainsi proposée rend le choix incertain à qui souhaite les intégrer au sein de son laboratoire. Aussi est-il essentiel de définir des critères de sélection.

Interface et architecture L'extrême majorité des LIMS disponibles est fondées sur une architecture à trois strates avec une API client très souvent web et à l'occasion JAVA, un service applicatif PHP ou JAVA et un SGBD très souvent relationnel ou moins fréquemment relationnel-objet à l'instar de GeneX qui utilise PostgresSQL. Cette architecture est à mon sens la plus adaptée à la vie d'un laboratoire. En effet, centralisées au sein d'un serveur, les données sont accessibles de plusieurs postes et n'ont pas besoin d'être synchronisées lorsque les travaux sont effectués sur plusieurs clients et peuvent bénéficier d'un système de sauvegarde performant. Les 2Hapi et RAD proposent une alternative en évitant une

Noms	Utilisation	Outils d'analyse	Format d'échange	Annotations actuallisées
2HAPI	en ligne	oui	html	non
BASE [152]	locale	oui	MAGE-ML	non
RAD	en ligne/locale	oui	MAGE-ML	non
GeneX [153]	locale	oui	MAGE-ML	non
MChips REFERENCE	locale	oui	XML	A voir

TAB. 2.2 – *Présentation de quelques LIMS et leurs caractéristiques*

installation locale car ils permettent le dépôt des données sur leur propre serveur. Cependant cette solution rend l'utilisateur dépendant des aléas inhérents à une connexion au serveur distant.

Outils d'analyse Tous les LIMS disponibles proposent des outils d'analyse statistique intégrés permettant d'agir directement sur les données qu'ils contiennent. Ces outils sont soit des logiciels à part entière soit des packages du langage statistique R tels que BioConductor. Ces outils sont indispensables et doivent proposer des méthodes de filtration, de normalisation et de comparaison. En sus, il est intéressant de proposer également des outils de regroupement et de classification.

Intégration des données hétérogènes Aucun des LIMS ne propose une actualisation des annotations des données. Ainsi, les annotations entrées dans la base sont figées. Il est pourtant intéressant de toujours avoir des données à jour, tant les annotations sont des données éphémères.

Format d'échange des données La quasi totalité des LIMS propose maintenant le format d'échange MAGE-ML. C'est aujourd'hui indispensable pour qui veut communiquer avec les autres bases de données.

Un effort tri-polaire (états-unien, européen, asiatique) est donc mené afin de permettre à chacun d'utiliser les données issues des technologies à haut débit d'analyse. Cet effort apporte des réponses quant aux moyens d'intégration des données et à la pertinence des détails à enregistrer. De plus il propose des standards en terme d'ontologie et de format d'échange de données.
Cependant en aval de ces contraintes contournées se dressent désormais d'autres interrogations concernant l'analyse statistique de ces données.

Chapitre 3

Analyses des données d'expression

Sommaire

3.1 Source et segmentation du signal

La quantification du signal à partir des images issues du scanner et l'attention portée à leur qualité sont les points clefs qui permettent de limiter les étapes de transformations (filtration, normalisation) et de rester au plus près de l'événement biologique.

3.1.1 Fluorescence et radioactivité

Les puces à ADN qui émettent en fluorescence nécessitent l'utilisation d'un instrument qui apporte une énergie excitatrice tel un laser et d'un système de détection tel le microscope confocal. Dans la pratique les fluorochromes, la cyanine 3-deoxyuridine triphosphate (Cy3) et la cyanine 5-deoxyuridine triphosphate (Cy5)[1] sont incorporés dans les échantillons par conjugaison à un uracile ou une cytosine. Il est important de noter à ce sujet que l'efficacité de l'incorporation des fluorochromes est inégale et que cette inégalité impose certaines particularités notamment lors de la définition du schéma expérimental (permutation des fluorochromes) et de la transformation des données (cf. page 87).

L'utilisation de la radioactivité est une autre méthode de marquage des acides nucléiques. Largement éprouvée dans les laboratoires de biologie moléculaire, elle est basée sur l'utilisation de l'isotope ^{33}P placé en position α ou γ du triphosphate nucléotidique. Les rayons X émis par les sondes déposées sur les puces de nylon impriment sur l'écran sensible une image. Bien que pour des raisons de risques sanitaires - cependant maitrisés - le marquage non-radioactif soit souvent préféré, le marquage radioactif offre l'avantage d'une plus grande sensibilité comparée à celle obtenue en fluorescence. Par ailleurs, les mesures restent linéaires sur un intervalle bien plus grand en radioactivité qu'en fluorescence. Enfin, l'utilisation des puces de nylon ont un coût moins élevé car il a été montré qu'ils pouvaient être utilisés à plusieurs reprises sans détérioration significative du signal [154].

3.1.2 Segmentation du signal

La quantification du signal émis par l'échantillon hybridé peut être affectée par une grande variété d'effets perturbateurs et générateurs de bruit-de-fond. Les sources de ces effets sont par exemple des altérations sur la position des spots, des formes et des contours irréguliers, une distribution inégale des sondes ADN au sein du spot, une qualité d'hybridation hétérogène ou encore plus simplement

[1] Cy3 est orange. Sa fréquence d'absorption maximale est 550nm et sa fréquence d'émission maximale est de 581nm. Les mêmes caractéristiques pour Cy5 sont respectivement vert, 649nm et 670nm.

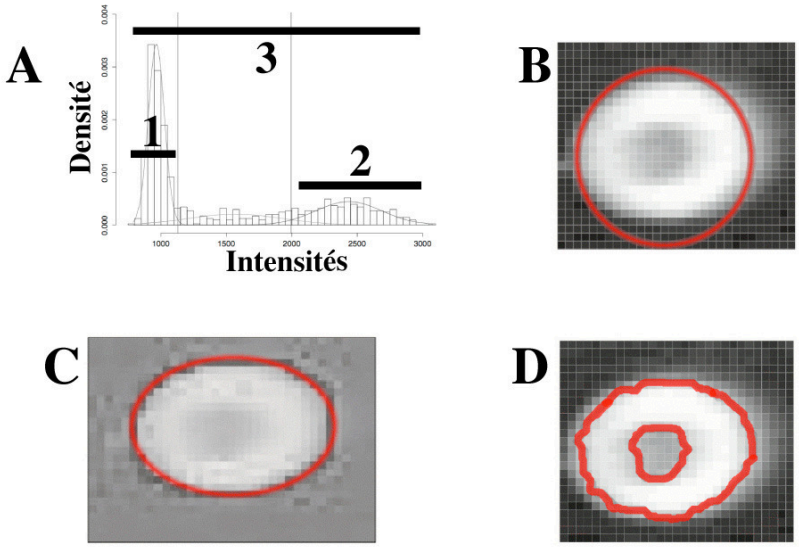


FIG. 3.1 – *Methodes de segmentation.* Il existe plusieurs méthodes de segmentations du signal chacune fondée sur un principe différent. (A) Une première est fondée sur la représentation en histogramme de la totalité des pixels des points (3). La définition d'une valeur limite permet de caractériser les sous-populations de pixels, à savoir bruit-de-fond (1) et signal propre (2). (B,C) Une seconde méthode repose sur l'application sur le point d'un cercle ou d'un ovale de diamètre fixe ou variable. Celui-ci delimite ainsi la frontière entre le bruit-de-fond et le signal. (D) La dernière méthode considère à la fois les valeurs des pixels et leur emplacement au sein des points.

des poussières et des précipités. La segmentation du signal, *ie* la détermination de la proportion de bruit de fond au sein du signal, permet donc idéalement de distinguer la part du signal dûe aux effets pertubateurs et celle dûe à l'information biologique.

Plusieurs logiciels tant académiques que commerciaux ont été développés pour quantifier et segmenter les signaux issus de la numérisation des puces à ADN. Trois méthodes sont actuellement utilisées [155] (cf. figure 3.1) :

- une segmentation basée sur l'intensité des pixels [156]. Un histogramme est calculé avec les intensités des pixels contenus dans un masque préalablement posé autour de chaque spot. Les valeurs d'intensités en deçà d'une valeur limite définie sont attribuées au bruit de fond. Le logiciel Quantarray [157] par exemple utilise cette méthode ;
- une segmentation spatiale. Un disque est posé sur chaque spot et les pixels situés hors du disque déterminent le bruit de fond. Des logiciels tels Scan-

Alyse [158], XDotsReader [159], GenePix et BlueFuse utilisent cet algorithme ;

– une segmentation mixte. Par exemple le module spotSegmentation [160] de la suite logicielle BioConductor [161] classe les pixels non pas en fonction d'un seule valeurs d'intensité limite mais par un algorithme de regroupement prenant en compte à la fois l'intensité des pixels et leur emplacement au sein du spot.

3.2 Schéma expérimental

À l'instar d'une démarche scientifique générale, avant d'engager une étude d'expression génique sur puces, il est important de décider d'une part quels gènes seront représentés sur la puce et d'autre part comment et combien d'échantillons seront déposés sur combien de puces. Prendre le temps de définir un bon schéma expérimental permet de quantifier les risques d'erreurs et les sources de variabilités, d'optimiser l'utilisation d'échantillons parfois précieux, de limiter ainsi les dépenses financières d'une technologie qui reste coûteuse et de pouvoir enfin s'assurer que les réponses obtenues répondront précisément en fin d'analyse aux interrogations initiatrices du projet [162].

3.2.1 Définitions

Le shéma expérimental des études d'expression sur puces est une structure à trois couches [163] (cf. figure 3.2) qui représente l'ensemble des sources de variations des mesures :

- La première représente les *unités expérimentales* comprenant d'une part les *réplicats biologiques* et les facteurs, *i.e.* les patients, animaux ou lignées cellulaires et leur traitement associé. Cette couche représente la variabilité inter-individus et représente à ce titre la plus importante source de variation expérimentale ;
- La seconde représente les *réplicats techniques*. Ce sont les échantillons issus des réplicats biologiques ;
- La troisième représente la variation au niveau des puces, conséquence d'une part des combinaisons de puces lors des hybridations et d'autre part des diverses caractéristiques des sondes.

Définir un bon schéma expérimental signifie donc déterminer la meilleure estimation des sources de variation qui contribuent réellement aux fluctuations des données afin de favoriser les effets biologiques au détriment des effets expérimentaux. L'erreur quadratique moyenne (*Mean Square Error*, MSE) est un bon indice de la qualité d'un schéma expérimental. Reprenant la structure générale définie précédemment (cf. figure 3.2), le MSE se compose de trois termes de variance correspondant aux trois couches

$$MSE = \sqrt{(\sigma^2_{exper} + \sigma^2_{techn} + \sigma^2_{sondes})}$$

Le calcul du degré de liberté du schéma expérimental permet une bonne estimation du MSE et la qualité d'un schéma expérimental peut être exprimée par son degré de liberté. Idéalement le nombre de degré de liberté ne doit pas être inférieur à 5 [163]. Afin d'augmenter ce nombre et de diminuer le MSE, une solution est l'augmentation de la taille de l'échantillon par des méthodes des réplicats.

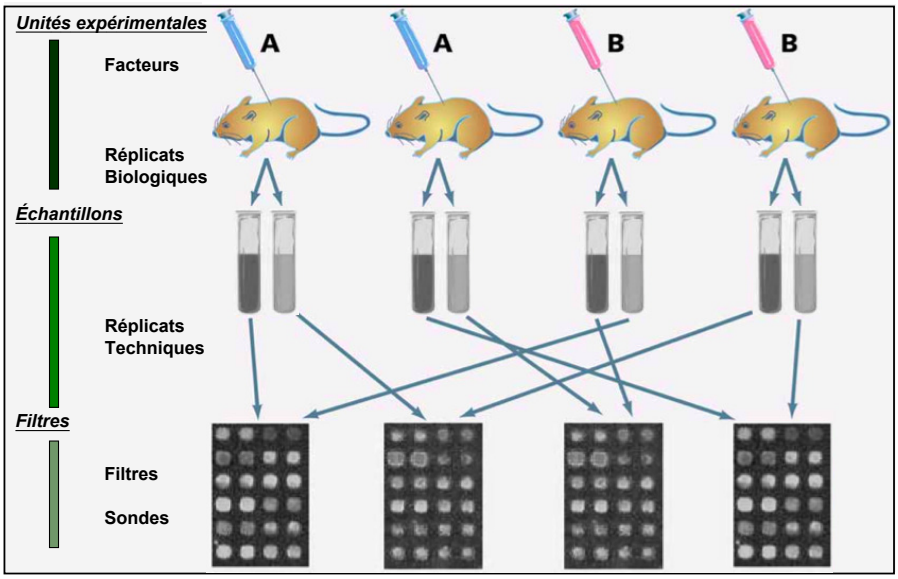


FIG. 3.2 – *Structure d'un schéma expérimental.* Le shéma expérimental d'une étude d'expression génique est une struture à trois couches qui décrit l'ensemble des variations des mesures. La première représente les *unités expérimentales* comprenant d'une part les *réplicats biologiques* et les facteurs, *i.e.* les patients, animaux ou lignées cellulaires et leur traitement associé. La seconde représente les *réplicats techniques*. Ce sont les échantillons issus des réplicats biologiques. La troisième représente les arrangements au niveau des puces. Elle décrit l'utilisation des puces et la nature et la position des sondes.

3.2.2 Réplications

Afin de diminuer l'influence des sources de variation non désirées, une solution est la moyennisation de celles-ci ; la moyenne fluctue moins que les variables qui la composent. Pour cette raison, la réplication tient une place importante dans le schéma expérimental car suivie d'une moyennisation, elle permet souvent une diminution du MSE.
Cependant le prix d'une augmentation de la précision des mesures est élévé. En effet la précision de l'estimation d'une variable augmente à la vitesse de $\sqrt{n}$, n étant le nombre de réplications.

3.2.2.1 Indépendance des mesures

Bien que certains admettent que trois réplications sont fréquemment suffisantes [164], définir la nature et le nombre de réplications dans schéma expérimental est fonction du degré d'indépendance des mesures répliquées. Cette notion est particulièrement importante car elle est la clé d'une optimisation des réplications qui

s'avèrent parfois être mal-appropriées et inutilement coûteuses. En effet si les mesures sont non-indépendantes, tout effet intrinsèque du paramètre sera également dupliqué. Aussi la moyennisation de ces mesures répliquées ne saurait réduire l'impact de cet effet sur la moyenne. La duplication de tels paramètres n'est donc pas intéressante [165]. Différentes techniques de réplication sont présentées.

3.2.2.2 Duplications des sondes

Très utile pour mesurer la qualité de l'hybridation, la duplication des sondes sur les puces est une méthode de réplication courante [166]. Il est cependant important de modérer ces propos dans la mesure où les sondes positionnées de manière adjacente sur la puce ne sont pas de réels duplicats. En effet leurs conditions expérimentales extrêment similaires et partagées rompent leur indépendance. Aussi afin de profiter des bénéfices de la moyennisation, il est péférable de disperser aléatoirement des réplicats sur l'ensemble de la grille de positions des sondes ce qui augmente leur degré d'indépendance.

3.2.2.3 Réplicats techniques

Les réplicats techniques sont issus d'une origine commune par un même processus d'extraction. Ainsi de fait, ces paramètres représentant les réplicats techniques sont des variables dépendantes qui ont une variabilité faible (moins élévée que la vraibilité biologique inter-individus). L'utilisation de réplicats techniques est donc utile pour corriger des biais technologiques mais certainement pas pour répondre à une question de biologie [167].

3.2.2.4 Unités expérimentales

La réplication des unités expérimentales est la source de variation la plus importante qui apparaît lors d'une étude d'expression génique. Il s'agit dans ce contexte d'échantillons d'organes ou de lignées cellulaires, extraits dans des conditions expériementales différentes et subissant un marquage dans des conditions également différentes. L'ensemble de ces considérations permet d'admettre que ces réplicats sont indépendants. Cette forme de réplication est en fait la plus appropriée et la moyennisation de ces mesures répliquées est particulièrement efficace [168].

3.2.3 Sélection des sondes

Particulièrement dans le cadre d'une étude sur l'humain, le grand nombre de gènes, l'existence d'introns et l'absence d'un séquençage complet et de qualité

rendaient l'amplification directe impossible. La solution était alors de s'orienter vers des petites portions de séquences exprimées *Expressed Sequence Tags*, EST) regroupées dans les bases de données publiques. Les ESTs sont ainsi la représentation de la portion transcrite du génome et les clones ADNc dont les ESTs sont dérivées sont devenues les premiers produits déposés sur les puces. Plusieurs outils de regroupements et d'annotation des ESTs sont disponibles tels UniGene [169], TIGR Gene [170]. Associée à une investigation bibliographique et à une recherche des *primers*, l'utilisation de ces outils permet d'élaborer une liste de clones ADNc qui seront déposés sur le puce.
Afin de faciliter la recherche des gènes interessants pour l'étude, il est pratique d'utiliser des outils informatiques de sélection, de rapatriement et de mise à jour automatique des annotations de sondes correspondants à ces gènes.

3.2.4 Représentation graphique

L'utilisation de graphes orientés est proposée par Yang&Speed [165] pour représenter les schémas expérimentaux des expériences d'expression géniques sur puces (cf. figure 3.3). Chacune des flèches représente une co-hybridation dans le cadre des puces à deux canaux d'émission et deux hybridations simples appariées (pour lesquelles l'expérimentateur s'appliquera à utiliser des conditions expérimentales les plus similaires possibles) dans le cas des puces à un seul signal d'émission. Il est alors possible de définir les effets à estimer par des chemins dont le nombre de noeuds rencontrés est inversement liée à la précision des estimations qu'ils représentent (cf. figure 3.3-2).

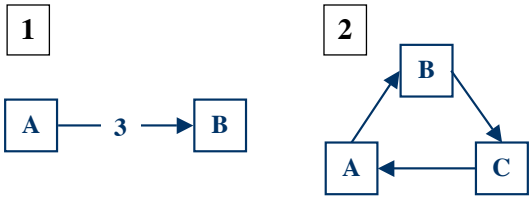


Fig. 3.3 – *Représentation graphique.* Chaque carré repésente un paramètre à étudier. Chaque flèche repésente une comparaison. (1) La comparaison directe de A et B est effectuée par trois réplications techniques. (2) A et B sont comparables suivants deux chemins : l'un direct de longueur 1 ($log(\frac{A}{B})$) et l'autre indirect de longueur 2 mais en conséquence moins précis ($log(\frac{A}{C}) - log(\frac{B}{C})$).

3.2.5 Types de schéma

Schéma avec référence Il s'agit du schéma expérimental le plus utilisé ; toutes les comparaisons directes sont effectuées face à un échantillon de référence (cf. figure 3.4-1). L'efficacité de ce schéma dépend beaucoup de la technologie employée. En effet dans le cadre des puces à deux canaux d'émission, ce schéma est pénalisé car la moitié des mesures est dissipée dans l'analyse de la référence [171, 163] ; ce qui n'est pas le cas pour les puces à un seul canal d'émission.
Par ailleurs le schéma avec référence est une approche satisfaisante du fait que le chemin séparant deux échantillons ne contient jamais plus qu'un noeud.
Enfin la définition de l'échantillon de référence est un point important dans l'élaboration du schéma expérimental. Une première approche consiste à utiliser un mélange d'ARNm dont la qualité requise est d'assurer que chaque sonde fournira un signal biologique. Une seconde approche consiste à utiliser un mélange des échantillons étudiés, ce qui diminue les différences absolues d'expression mais permet d'améliorer la normalisation d'échantillons qui sont parfois très différents [172].

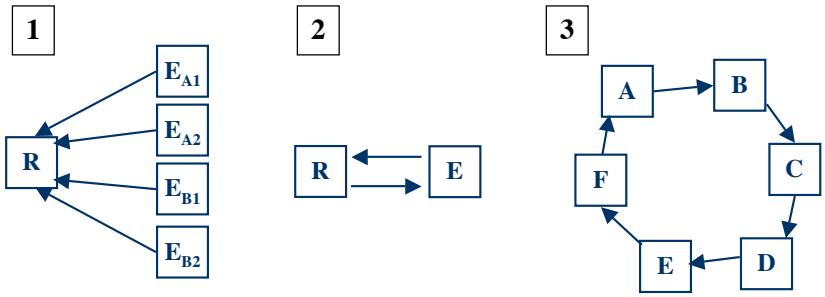


FIG. 3.4 – *Types de schéma expérimentaux* (1) Schéma avec référence. Toutes les comparaisons directes sont effectuées face à un échantillon de référence. (2) Schéma à comparaison directe. Les échantillons sont étudiés en comparaison directe limitant ainsi les sources de variations. (3) Schéma circulaire. Les échantillons sont comparés les uns aux autres dans une chaîne ordonnée circulaire

Schéma à comparaison directe Ce type de comparaison permet de limiter les sources de variation du signal dûes au puce (et ce d'autant plus lorsque la technologie utilisée permet une double hybridation simultanée) (cf. figure 3.4-2). La comparaison deux-à-deux des échantillons n'est cependant pas toujours réalisable et il est alors important de regrouper sur un même puce les comparaisons de plus grand intérêt.

Schéma en boucle Le schéma en boucle (cf. figure 3.4-3) dans lequel les échantillons sont comparés les uns aux autres dans une chaîne ordonnée circulaire [171] est une alternative au schéma avec référence. Il est cependant mal adapté aux expériences menées sur un grand nombre d'échantillons.

La recherche sur le thème des schémas expérimentaux optimisés est très active. En effet, fortes d'une reproductibilité des résultats toujours accrue (bien que souvent issus de plateformes technologiques différentes) [173], la compléxité des études comparatives est grandissante et les méta-analyses intégratives utilisants les données de plusieurs centaines d'expériences emmergent progressivement [174]. Ainsi les nouveaux schémas expérimentaux ont-ils obligation de refléter cette compléxité et de mettre à disposition de l'expérimentateur des outils mathematiques pour apporter les réponses aux questions initiales posées.

3.3 Modélisation des données

3.3.1 Régression linéaire multiple

Le principal intérêt de la régression linéaire multiple est de estimer la relation entre plusieurs variables indépendantes (*facteurs*) et une variable dépendante, notamment la relation entre des traitements ou des pathologies et le niveau d'expression d'un gène. Cette estimation est rendu par la minimisation de la différence entre les valeurs estimées et les valeurs observées grâce par exemple à la méthode des moindres carrés (cf. figure 3.5-A) :

$$min\left(\sum_{i=1}^{n}\epsilon_i\right)$$

Supposons par exemple la relation à deux facteurs entre l'estimation du niveau d'expression ($\hat{y}$) d'un gène au cours de deux traitements. Celle-ci peut être représentée par l'équation de regression suivante :

$$\hat{y} = \beta_0 + \beta_1.Traitement_1 + \beta_2.Traitement_2$$

β_1 et β_2 sont les coefficients et représentent respectivement la contribution des facteurs *traitements* 1 *et* 2 dans l'estimation de la valeur du signal (cf. figure 3.5-B).

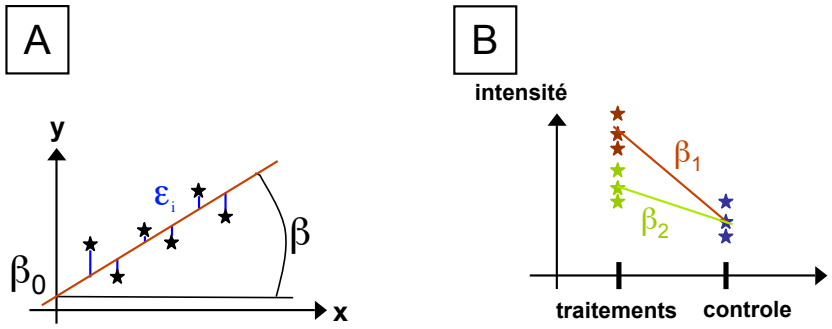


FIG. 3.5 – *Regression lineaire.* A. Régression linéaire simple. En bleu sont représentées les distances ϵ_i entres les valeurs y observées et les valeurs $\hat{y}$ estimées par la méthode des moindre carrés. La droite en brun est la représentation de l'équation $\hat{y} = \beta_0 + \beta x$, dont β est le coefficient directeur et β_0 est le terme d'intersection. B. Régression linéaire multiple. Les points en vert et brun représentent les valeurs expérimentales de deux traitements à comparer aux valeurs expérimentales d'un contrôle. Pour chaque traitement le coefficient directeur de la droite de régression linéaire peut-être estimé par la méthode des moindre carrés.

Cette notation est extensible à k facteurs :

$$\hat{y} = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + ... + \beta_k.x_k$$

3.3.1.1 Sources de variations

Dans la pratique nombre de sources de variations interviennent et influent sur la valeur du signal d'intensité. Ces sources sont par exemple le temps au cours d'un processus biologique, les différents types de tissus ou encore les différents traitements apportés aux échantillons.
Il a par ailleurs été montré que la séquence des sondes déposées sur la puce et en conséquence l'affinité de la sonde envers sa cible, était également une source de variabilité du signal.
Enfin comme une expérience repose souvent sur l'utilisation de plusieurs puces à ADN et utilise parfois plusieurs types de marqueurs, les puces utilisées et les marqueurs utilisés sont alors également des sources de variabilité du signal.
La valeur d'intensité mesurée est donc la resultante d'une relation complexe de l'ensemble de ces sources ou facteurs.

Ainsi quatre facteurs majeurs sont identifiés :
- les variations biologiques sujets de l'étude ;
- les gènes ;
- les puces à ADN ;
- le marquage.

De ces quatre facteurs, $2^4 = 16$ effets expérimentaux sont déduits. Explicitement il y a d'abord quatre effets de base : les variations biologiques (ζ), les marquages (δ), les puces (α) et les gènes (γ). Il existe de plus des interactions entre ces sources ; six interactions binaires, quatre tertiaires et une quaternaire [175].
Parmi les interactions binaires, l'interaction $\zeta \times \gamma$ représente la différence d'expression des combinaisons $facteur \times gène$ non expliqués par les effets moyens des facteurs et des gènes. En conséquence, identifier les gènes dont l'expression varie en fonction des différents facteurs revient à identifier les différences non-nulles de $\zeta \times \gamma$.

3.3.1.2 Modèle linéaire

La modélisation linéaire des données est une extension de la régression linéaire multiple grâce à laquelle il est alors possible d'utiliser non plus n observations d'une seule variable $y_{1..n}$ mais n observations de m variables $y_{1..n,1..m}$.
Le modèle linéaire le plus simple est un modèle global, *i.e.* que ses paramètres sont déterminés sur l'ensemble des gènes. Il repose par ailleurs sur l'hypothèse que les quatres principaux facteurs de variabilité définis précédemment sont des

constantes uniformément appliquées aux données.
Il est important de noter alors que parcequ'elles sont difficilement explicables par un processus physique particulier et par souci de simplification, certaines interactions dont notamment celles de haut niveau telles les interactions tertiaires et quaternaires sont volontairement omies dans les équations suivantes [176] :

$$y_{i,j,k,l} = \mu + \alpha_i + \delta_j + \zeta_k + \gamma_l + (\alpha_i\gamma_l) + (\zeta_k\gamma_l) + \epsilon_{i,j,k,l}$$

$$\begin{array}{lll} avec: & y_{i,j,k,l} : mesure & \alpha_i : effet\ du\ puce \\ & i : puce & \delta_j : effet\ du\ marquage \\ & j : marquage & \zeta_k : effet\ du\ traitement \\ & k : traitement & \gamma_l : effet\ du\ gène \\ & l : gène & (\alpha\gamma)_{il} : effet\ gène \times puce\ (spot) \\ & \mu : ratio\ globa & (\zeta\gamma)_{kl} : effet\ gène \times traitement \\ & & \epsilon_{i,j,k,l} : erreur \end{array}$$

L'attractive simplicité de ce modèle en est également le point faible. En effet, lorsque le nombre de gènes devient considérable comme fréquemment pour les études d'expression génique, l'estimation précise des paramètres du modèle fondée sur un ensemble de données dont la variabilité peut parfois être importante peut s'avérer difficile. Il a cependant l'intérêt de poser les fondements pour l'élaboration de modèles plus complexes mieux adaptés.

Aussi comme la solution repose dans la complexification du modèle linéaire initial, Kerr propose le modèle dit ANOVA qui est une décomposition du modèle précédent [175, 177] :

- le premier, global regroupant les paramètres appliqués uniformément à l'ensemble des données ;
- le second, regroupant les paramètres particulier de chaque gène.

$$y_{i,j,k,l} = \underbrace{\mu + \alpha_i + \delta_j + \zeta_k}_{terme\ global} + \underbrace{W_{i,j,k}}_{\substack{terme\ particulier \\ pour\ chaque\ gène}}$$

Cette modélisation linéaire des données d'expression ne prétend pas être exhaustive tant le concept est adaptatif. Il est ainsi envisageable d'intégrer au modèle des paramètres de variabilité biologique, de réplicats techniques ou encore de réplicats biologiques [178].

À l'évidence l'intérêt n'est pas de complexifier le modèle au maximum. En effet la complexification maximale conduirait à un modèle dont le nombre de termes

correspond au nombre de données. Un tel modèle représente certes le mieux possible le comportement des données mais n'apporte aucun intérêt tant sa complexité devient encombrante lors de sa manipulation avec les outils mathématiques. La solution est donc un modèle de niveau de complexité intermédiaire qui permet de prédire sans trop d'erreur le comportement des données.

3.3.2 Représentation matricielle

3.3.2.1 Matrice d'expression

Il est possible de représenter l'expression des gènes grâce à une matrice d'expression Y de dimension $n \times m$

$$Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1m} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & y_{n3} & \cdots & y_{nm} \end{pmatrix}$$

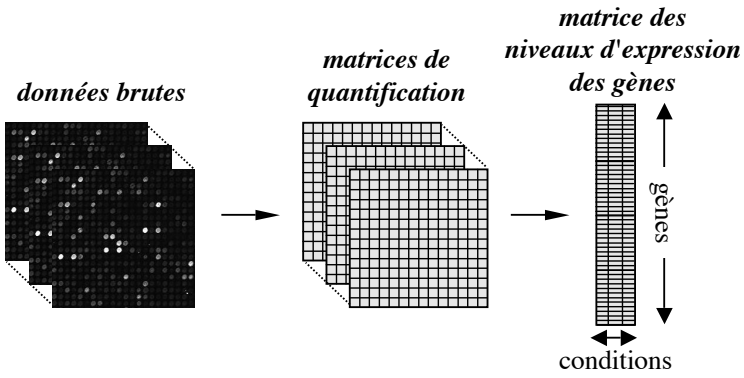


FIG. 3.6 – *Matrice d'expression.* La lecture des images par le *scanner* permet l'obtention de *matrices de quantification* dont les valeurs correspondent à chacune des sondes sur le puce pour une hybridation donnée. Le regroupement de ces matrices conduit à la composition d'une *matrice d'expression* pour laquelle les colonnes représentent les hybridations et les lignes les sondes.

Ainsi chacune des m conditions expérimentales (ou échantillons) est une colonne dans la matrice des niveaux d'expression et chacun des n gènes est une ligne de cette matrice (cf. figure 3.6).

3.3.2.2 Matrice du schéma expérimental

Considérons désormais une expérience suivant un schéma expérimental avec 3 traitements A, B, C chacun comparé à une référence R. Pour chaque gène i, seuls les trois coefficients correspondants aux trois traitements sont alors considérer par souci de simplification :

$$\begin{aligned}\beta_{iA} &= \mu_{iA} - \mu_{iR}\\ \beta_{iB} &= \mu_{iB} - \mu_{iR}\\ \beta_{iC} &= \mu_{iC} - \mu_{iR}\end{aligned}$$

Le modèle linéaire d'une telle expérience peut être formulé de la manière suivante :

$$\begin{aligned}Y_{iA} &= 1.\beta_{iA} + 0.\beta_{iB} + 0.\beta_{iC} + e_{iA}\\ Y_{iB} &= 0.\beta_{iA} + 1.\beta_{iB} + 0.\beta_{iC} + e_{iB}\\ Y_{iC} &= 0.\beta_{iA} + 0.\beta_{iB} + 1.\beta_{iC} + e_{iC}\end{aligned}$$

L'ensemble de ces trois équations est similaire à l'équation suivante :

$$Y = \begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}.\begin{pmatrix}\beta_{iA}\\ \beta_{iB}\\ \beta_{iC}\end{pmatrix} + \begin{pmatrix}\epsilon_A\\ \epsilon_B\\ \epsilon_C\end{pmatrix}$$

Soit pour l'ensemble des gènes,

$$Y = S.b + \epsilon$$

Y *les estimations des valeurs d'expression*
S *matrice du schéma expérimental*
b *vecteur des coef ficients*
ϵ *vecteur des résidus.*

Les coefficients représentent la contribution des facteurs à la variabilité du signal d'intensité. Cependant l'expérimentateur est souvent intéressé dans la comparaison de deux facteurs. Pour cela il faut définir des *contrastes*. Par exemple afin de déterminer la difference entre les traitements A et B, on définit le contraste $\mu_{iA} - \mu_{iB} = \beta_{iA} - \beta_{iB}$, représenté par la matrice c :

$$\begin{aligned}\beta_{iA} - \beta_{iB} &= \begin{pmatrix} 1 & -1 & 0\end{pmatrix}.\begin{pmatrix}\beta_{iA}\\ \beta_{iB}\\ \beta_{iC}\end{pmatrix}\\ &= c.b\end{aligned}$$

Et de déterminer ainsi la contribution de ce contraste dans la variation du signal de chacun des gènes i.

Parmi les logiciels qui fondent leur analyse des données d'expression sur la modélisation linéarisation des données d'expression, LIMMA (LInear Models for MicroArray) [179] associé à la suite logicielle BioConductor est l'un des plus utilisés [180].

3.4 Transformation des données

3.4.1 Filtration

La variabilité accrue des intensités faibles proches du bruit de fond impose une filtration de ces valeurs potentiellement faussement positive. Une méthode simple et fréquente consiste à n'utiliser pour l'analyse que les valeurs significativement différentes du bruit de fond.
La segmentation du signal permet de mesurer le bruit de fond global ou local pour chaque spot s sur la puce. Considèrant que le bruit de fond est distribué selon une loi Normale, il est possible de calculer alors sa variance afin de définir quels sont les spots qui apportent un réel signal biologique et quels sont ceux dont le signal est trop faible et sont confondus alors dans le bruit de fond. Une valeur limite permet alors la séparation des deux catégories de spots au risque α d'erreur [181].

$$I_s^{signal} > 2.\sigma[I_{1..s}^{bruit\ de\ fond}] > I_s^{bruit\ de\ fond}, \alpha = 0,05$$

Bien que différentes approches existent [2], cette simple méthode permet d'augmenter considérablement la qualité des analyses ultérieures [172].

3.4.2 Normalisation

L'objet de la normalisation est de diminuer l'apport au signal de chaque effet induit par la technologie afin de mettre en lumière les effets biologiques étudiés. Ces effets inhérents à la technologie qui parasitent l'information biologique proviennent du robot-spotteur, de l'inégale efficacité d'incorporation des fluorochromes (pour les technologies à deux couleurs), du processus d'hybridation et plus généralement des conditions environnementales. Plus les effets technologiques seront reduits et plus la comparaison des niveaux d'expression entre l'échantillon et la référence sera proche de la réalité biologique.

3.4.2.1 Hypothèses

L'hypothèse primordiale considère que les cellules ont globalement toutes une même capacité de production ARNm. En d'autre termes, l'augmentation du niveau d'expression de certains gènes est compensée par la diminution de celui d'autres gènes. Ainsi globalement, l'expression de la grande majorité des gènes représentés sur la puce ne varie pas entre l'échantillon et la référence.

[2] http ://www.bioconductor.org/packages/bioc/1.7/src/contrib/html/genefilter.html

Ce postulat se traduit mathématiquement par le fait que la moyenne des ratios échantillon *vs* référence tend vers 1. À cette fin, différentes méthodes ont été décrites pour déterminer le facteur de normalisation compensateur des effets technologiques afin de re-centrer les quotients des intensités sur 1.

3.4.2.2 Transformations logarithmiques

La transformation logarithmique des quotients des intensités (*log-ratios*) apporte au moins trois avantages. Elle permet d'abord de passer d'un modèle multiplicatif à un modèle additif. Ainsi, équations exponentielles et autres facteurs deviennent droites et constantes à additionner.
De plus, alors que dans une représentation linéaire les ratios sont comprimés dans l'intervalle $[0; 1]$, la représentation logarithmique les répartit de façon symétrique autour d'une valeur centrale qui n'est plus 1 mais 0. Ainsi une diminution du niveau de l'expression est représentée par un valeur négative et une augmentation du niveau de l'expression par une valeur positive (cf. figure 3.7). Enfin, la transformation logarithmique des ratios permet l'utilisation de tests statistiques paramétriques dans la mesure où la variance est stabilisée et la distribution des log-ratios peut-être apparentée à une distribution de loi Normale.

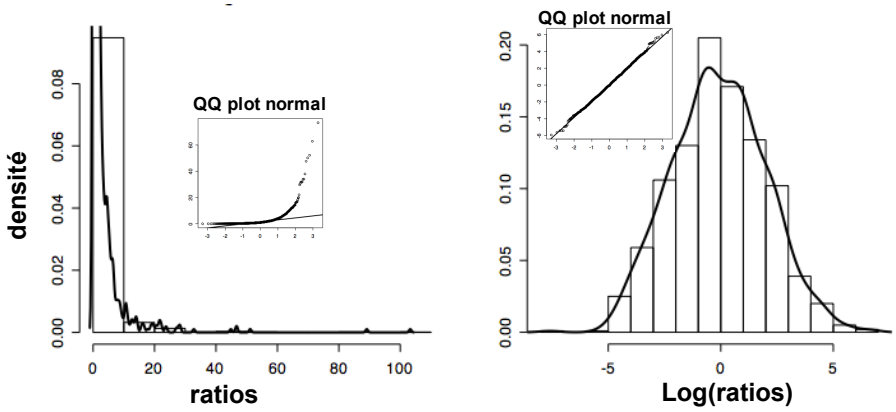


FIG. 3.7 – *Transformation logarithmique.* La transformation logarithmique des données permet d'une part le centrage des valeurs sur 0 et d'autre part d'établir une répartition symétrique autour de la valeur centrale. Le graphe *QQ plot normal* représente la projection des quantiles observés contre les quantiles d'une distribution de loi normale. La diagonale (droite de HENRY) indique que les données observées suivent une distribution de loi Normale. Ainsi après transformation logarithmique, la distribution des données d'intensités suit une loi Normale.

3.4.2.3 Mise à l'échelle

Une spécificité des technologies à un seul canal d'émission est le fait que les échantillons et les références ne sont pas analysés en même temps et qu'ils sont de plus souvent déposés sur des puces différentes. En conséquence il peut exister une différence d'échelle dans les distributions des intensités (cf. figure 3.8) affectant alors le calculs des ratios. Des méthodes de normalisations notamment celle fondée sur la représentation quantile-quantile des données et celle fondée sur une analyse de la variance ont alors été développées et présentée ci-après.

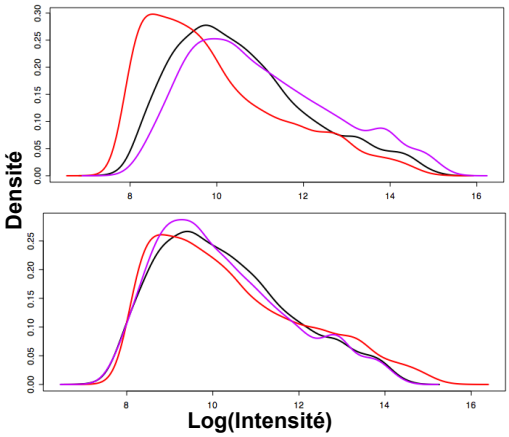


FIG. 3.8 – *Distribution des intensité.* Les distributions des valeurs des intensités sont fréquemment différentes notamment en ce qui concerne leur variance. Une mise à l'échelle des valeurs permet alors de rendre les variances des distributions plus semblables. Après transformation les distributions sont désormais comparables.

Quantiles-Quantiles Le principe de la normalisation par quantiles [182] est de rendre les distributions des intensités identiques pour chaque sonde sur l'ensemble des puces. La démarche est motivée par l'idée que sur un graphique quantile-quantile (Q-Q)[3], la projection P des quantiles de deux distributions identiques affiche une diagonale d de coefficient 1 et d'interception 0, dite droite de HENRY. Lorsque les distributions sont différentes (cf. figure 3.9) la projection s'éloigne sensiblement de cette diagonale. La normalisation Q-Q étend le concept à N distributions projetées dans N dimensions. La diagonale d'identité est alors le vecteur de coordonnées $d_N = (\frac{1}{\sqrt{N}}, ..., \frac{1}{\sqrt{N}})$. La normalisation Q-Q consiste donc

[3]Un graphique QQ est un graphique des quantiles d'un premier jeu de données contre ceux d'une second jeux de données. Un quantile est la fraction de points inférieurs à une valeur seuille donnée.

en l'estimation et la modification des paramètres des N distributions afin de rendre ces dernières identiques.
La limite de cette méthode est sa rigidité. En effet elle impose l'égalité aux valeurs des quantiles. Ainsi arrive-t-il parfois qu'aux valeurs de queue de distribution leur soient assignée une même valeur sur l'ensemble des puces.

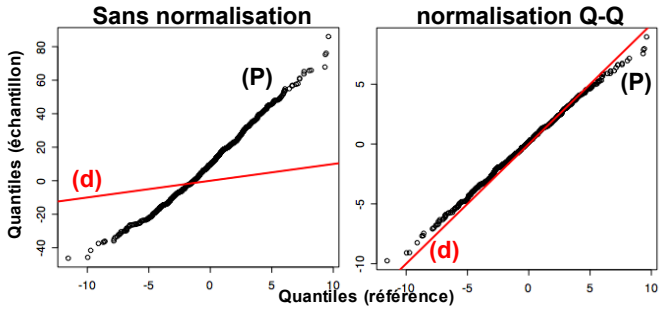


FIG. 3.9 – *Normalisation Quantile-Quantile.* La normalisation Quantile-Quantile consiste à transformer les données afin de rendre les quantiles de chaque distribution identiques. Un graphe Quantile-Quantile permet la représentation de deux distributions. Lorsque les distributions sont identiques, les quantiles sont identiques et les valeurs se projettent (P) alors sur la diagonale d (droite de HENRY). La normalisation Q-Q étend le concept à N distributions dans N dimensions.

Analyse de la variance Reprenant le concept de modélisation linéaire des valeurs d'expression (cf. page 77), l'ensemble des sources des variations qui affectent les signaux sont mathématiquement traduits dans l'équation suivante

$$Y_{i,j,k,l} = \underbrace{\mu + \alpha_i + \delta_j + \zeta_k}_{\substack{facteur\ de \\ normalisation\ globale\ (c)}} + \underbrace{W_{i,j,k}}_{\substack{paramètres \\ d'intéraction \\ pour\ chaque\ gène}}$$

$$avec : \quad \hat{W} = \gamma_l + (\alpha\gamma)_{il} + (\zeta\gamma)_{kl} + \epsilon_{i,j,k,l}$$

Cette équation est constituée de deux termes : le premier est un ensemble de paramètres représentant le facteur de normalisation globale c et le second représente les paramètres d'interactions propres à chaque gène. Il est possible d'estimer la valeur de la constante c par une analyse de la variance (*anova*) des intensités [176, 183], *i.e.* estimer quelle est la part de responsabilité de chaque terme dans la variation des données.

3.4.2.4 Facteur de normalisation comme constante

Moyenne et médiane La plus simple des normalisations considère le facteur de normalisation comme une constante. Après transformation dans une échelle logarithmique, la constante c est soustraite de chacun des log-ratios (cf. figure 3.10). Les log-ratios sont ainsi en moyenne centrés sur 0. La moyenne et la médiane des log-ratios sont deux estimations simples de cette constante. La médiane possède l'avantage sur la moyenne d'être moins sensible aux valeurs extrêmes et en conséquence de mieux représenter la distribution des log-ratios.

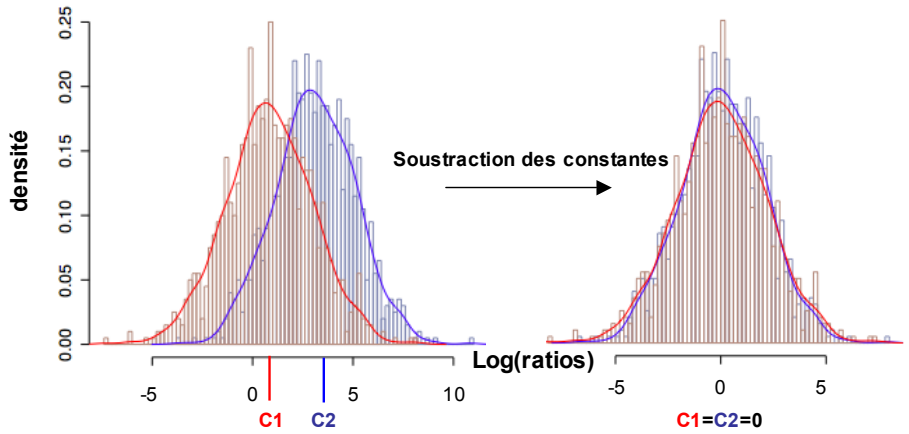


FIG. 3.10 – *Normalisation globale.* Le centrage des distributions par soustraction d'une constante c est une méthode courament utilisée lors des transformations des données avant d'entamer une analyse statistique. c peut être par exemple soit la moyenne soit la médiane de la distribution. Après soustractions des constantes (c_1 et c_2) à chacune des valeurs des distributions respectives, les deux distributions sont centrées sur 0.

Régression linéaire Reprenant l'idée que pour chaque gène, son niveau d'expression dans l'échantillon est une fonction linéaire de celui dans la référence, il est possible de normaliser les données en estimant les paramètres de la droite de régression par la méthode des moindres carrés et d'en déduire la valeur de c afin de déplacer cette droite vers une pente de valeur 1 et une interception de valeur 0 [184, 121].
Reprenant donc l'équation (1) (cf. page 76), la meilleure estimation de β_1

$$\hat{\beta}_1 = \frac{\sum_{l=1}^{n}(S_l - \overline{S})(R_l - \overline{R})}{\sum_{l=1}^{n}(R_l - \overline{R})^2} \ (\textit{pour chaque gène } l)$$

Puis l'estimation de β_0 peut se calculer ainsi

$$\hat{\beta}_0 = \overline{S} - \hat{\beta}_1.\overline{R}$$

Enfin la mesure normalisée des intensités vaut donc pour chaque gène l

$$R' = \left(\frac{G_l - \hat{\beta}_0}{\hat{\beta}_1}\right) \quad et \quad S'_l = S_l$$

Statistiques des ratios Une autre approche de normalisation est la méthode définie par CHEN [156] et basée sur les statistiques des log-ratios. Considèrant que bien que pour certains gènes le niveau d'expression puisse varier, dans des cellules semblables, la quantité totale d'ARNm est la même dans tous les échantillons. En conséquence il est admis l'hypothèse que ces gènes suivent une distribution normale de même moyenne et de même variance indépendament des échantillons. Fondée sur un algoritme E-M [4] d'estimation du maximum de vraissemblance, la méthode itérative ainsi développée permet l'estimation des paramètres de cette distribution, de la constante c ainsi que l'identification de gènes dont la variation du niveau d'expression est significative. Par la suite la méthode fut améliorée notamment du point de vue de la prise en compte des valeurs proches du bruit de fond [185]

3.4.2.5 Facteur de normalisation comme fonction

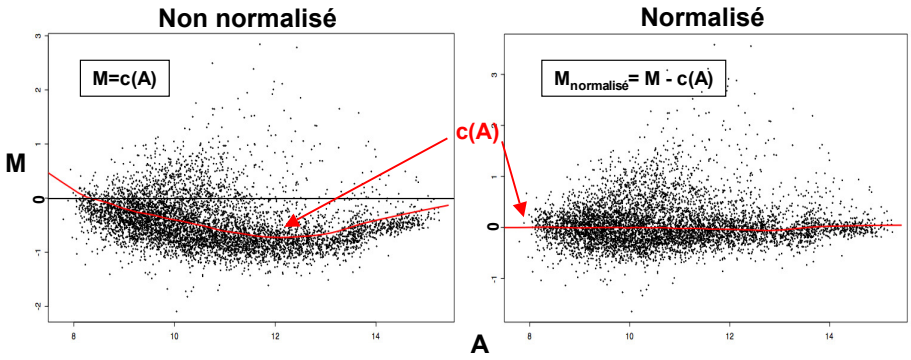


FIG. 3.11 – *Fonction de normalisation.* Les log-ratios (M) sont dépendants des log-intensités moyennes (A) et cette dépendance n'est pas linéaire. Aussi les valeurs des M de nécessitent-elles pas toutes le même facteur de normalisation. La fonction c(A) est la fonction de normalisation qui permet de déterminer les log-ratios normalisés ($M_{normalisé}$) en fonction des A.

[4] Algorithme itératif d'estimation de variables par *l'estimation du maximum de vraisemblance* . L'estimation du maximum de vraisemblance (EMV) est une méthode statistique dont l'objet est de déterminer un ensemble d'estimations de paramètres telle que la probabilité d'obtenir l'échantillon est maximale. L'EMV est recalculée à chaque itération jusqu'á l'obtention des meilleures probabilités

Il est démontré par ailleurs que les log-ratios sont dépendants des intensités et que de plus, cette dépendance n'est pas linéaire en particulier pour les intensités faibles [186]. En d'autres termes, les log-ratios des intensités faibles ne necessitent pas les mêmes facteurs de normalisation que ceux des intensités élevées.
La conséquence d'une telle observation est que c n'est plus une constante mais une fonction de l'intensité (cf. figure 3.11). Notons les paramètres suivants :

$$\begin{aligned} & M = log\left(\frac{S}{R}\right) && le\ log-ratio \\ & A = \frac{1}{2}.log(S.R) && la\ log-intensité\ moyenne \\ & alors && \hat{M} = \hat{c}(A) \\ & && M_{normalisé} = M - \hat{M} \end{aligned}$$

Plusieurs méthodes ont été proposées pour estimer la fonction de normalisation $c(A)$.

Méthode *Lowess* Les méthodes pour déterminer la fonction $\hat{c}(A)$ sont notamment fondées sur des regressions linéaires, certaines robustes mais stringentes [187] et d'autres plus souples telle Lowess (*Locally WEighted Scatterplot Smoothing*) [188]. Cette méthode est une méthode de lissage des données fondée sur une régression linéaire pondérée localement[5].
Fréquemment appliquée aux données d'expression [189, 190, 191] et particulièrement efficace sur les fortes sources de variations telle la différence d'incorporation des fluorochromes, cette méthode itérative fonctionne sur le schéma suivant. Centrée sur chaque valeur de M à corriger noté M_c, une fenêtre d'analyse est définie. Les coefficients d'une regression linéaire pondérée sont alors calculés pour cette fenêtre. La pondération des coefficients est déterminée pour chacun des M voisins en fonction de sa distance croissante au M_c. Les M situés à l'extérieur de la fenêtre ont un poids nul. La procédure terminée, la fenêtre se déplace le long des axes des A et chacune des valeurs de M subit à son tour une regression linéaire pondérée (cf. figure 3.12 A).

3.4.2.6 Normalisation globale et locale

Beaucoup des méthodes citées précédemment peuvent être appliquées soit globalement soit localement. En effet, en particulier l'utilisation d'aiguilles lors du dépôt des sondes mais également les conditions d'hybridation parfois différentes

[5]Loess est une extension de la méthode Lowess. Leur différence réside dans le modèle de linéarisation utilisé. Alors que Lowess est fondée sur un modèle polynomial *linéaire*, Loess est fondé sur un modèle polynomial *quadratique*.

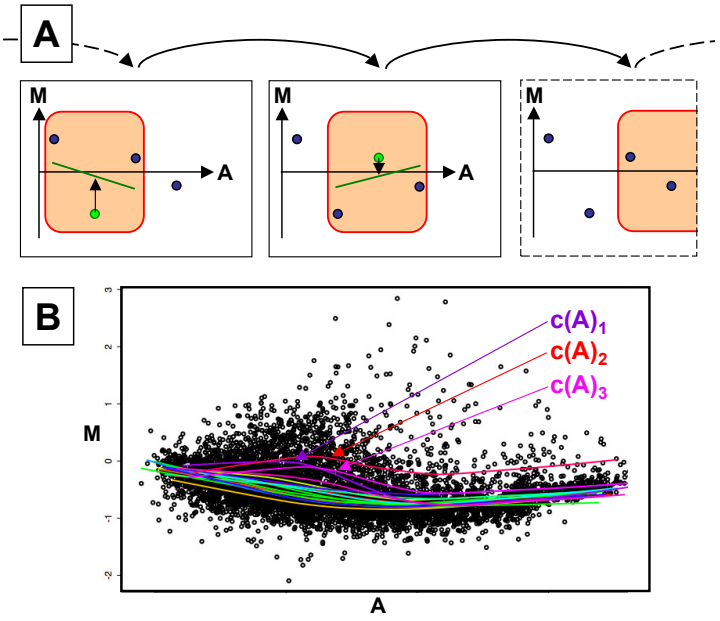


FIG. 3.12 – *Méthode Lowess.* (A) Lowess est une méthode de lissage des données fondée sur une régression linéare pondérée localement. Centrée sur chaque valeur de M à corriger noté M_c, une fenêtre d'analyse est définie. Chaque M_c est estimé de nouveau par calcul de régression linéaire pondérée. Le poids est atténué pour chacun des M voisins en fonction de sa distance croissante au M_c. (B) Cette méthode peut également être appliquée lorsque les point regroupés par exemple en fonction de l'aiguille de dépôt. Pour chaque groupe g une fonction $c(A)_g$ est estimée. L'estimation des fonctions $c(A)_{1,2,3}$ montrent nettement un effet imputable aux aiguilles de dépôts $1, 2et3$.

à la surface du puce font que l'emplacement de la sonde sur le puce est une source importante de variabilité du signal [191].

Fréquemment utilisé dans les études utilisant les puces à ADN, une adaptation particulière de Lowess intitulée *printtips Lowess* tient compte de la localisation des points sur le puce et de l'effet physique de l'aiguille qui les a déposé. Les points sont ainsi regroupés en g groupes. Une fonction $c(A)_g$ est alors estimée pour chacun des groupes g de points (cf. figure 3.12 B).

3.5 Analyse des données

3.5.1 Différences d'expression

L'intérêt d'une étude sur le niveau d'expression des gènes est à l'évidence de découvrir les gènes dont le niveau d'expression varie entre l'état de référence et celui testé. Deux types d'approches sont généralement proposés bien que la seconde deviennent prépondérante. La première regroupe des méthodes dont la détermination de la différence d'expression de chaque gène est fondée sur l'intégration des données de l'ensemble des gènes. La seconde regroupe les méthodes dont l'analyse est effectuée de façon particulière sur chacun des gènes.

3.5.1.1 Méthodes intégrales

Seuil global Les premières études d'expression génique portées par les puces d'expression des gènes triaient les gènes grâce à un facteur seuil de changement de niveau d'expression fixé pour l'ensemble des gènes. Il était empiriquement admis qu'une valeur de 2 était suffisante. Ainsi un niveau d'expression qui variait de plus de deux fois entre l'échantillon et la référence était significatif.
Cependant cette approche ne considère aucunement le fait que la variabilité des intensités faibles est plus importante que celle des intensités élevées [186]. Elle induit alors des faux-positifs aux faibles intensités et des faux-négatifs aux fortes intensités. Une approche plus appropriée est l'utilisation de seuils non constants adaptés à la fonction de variabilité des intensités.

Z-score Une seconde approche consiste à calculer un Z-score sur l'ensemble des log-ratios (cf. figure 3.13 A) et de déterminer ainsi les ratios anormaux avec un niveau de risque d'erreur défini.
Afin de tenir compte de la fonction de variabilité des intensités, il est possible de définir une fenêtre glissant le long des données au sein de laquelle est calculé le Z-score pour les valeurs de cette fenêtre [192]. Ainsi le Z-score calculé est une fonction de la variabilité des intensités.
Notant $\sigma^f_{log_2(T_i)}$ l'écart-type des log-ratios de la fenêtre f pour le gène i

$$Z_i^f = \frac{log_2(T_i)}{\sigma^f_{log_2(T_i)}}$$

cela nous permet de considérer que la différence d'expression pour le gène i est significative au risque $\alpha = 0.05$ si $|Z_i^f| > 2$.

Une méthode dérivée du calcul du Z-score local propose la détermination d'une fonction de lissage des Z-scores. Cette fonction prend alors en compte la

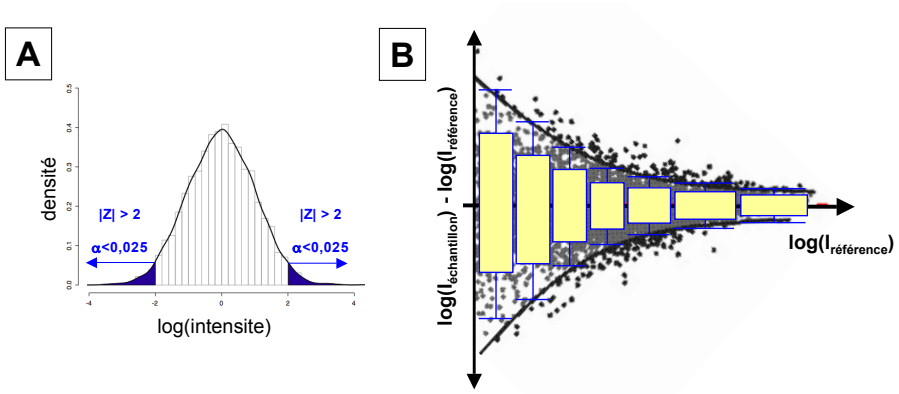


FIG. 3.13 – *Différence d'expression.* (A) Il est possible de déterminer avec un risque d'erreur donné ($\alpha < 0.05$) les valeurs de ratios anormales en calculant un Z-score pour chacune des valeurs. Ainsi à tout $|Z| > 2$ correspondent des variations significatives du signal. (B) Une extension de cette approche consiste à définir des fenêtres au sein desquelles sont calculés des Z-scores. Dans un second temps, chaque valeur de Z-score est ajustée en fonction des autres grâce à la détermination d'une fonction de lissage.

variabilité locale des intensités tout en lissant les valeurs de Z-score sur l'ensemble de données (cf. figure 3.13 B) [193].
La limite principale des approches par Z-score est leur caractérisique canonique. En effet ces méthodes détermineront sytématiquement des gènes différentiellement exprimés notamment lorsqu'aucune différence d'expression est attendue lors par exemple de compararisons d'échantillons identiques situés en conditions expérimentales identiques.

3.5.1.2 Méthodes partielles

Statistiques *t* Il est possible pour chaque gène de calculer une statistique t sur l'ensemble des réplicats d'échantillons et de référence.

$$t_g = \frac{\bar{y_g}}{\sigma_g \sqrt{\frac{1}{n_g}}}$$

L'hypothèse $H0$ est alors que le niveau d'expression moyen entre les échantillons et la référence n'est pas différent. La principale limite de cette approche est portée sur l'estimation de T qui devient douteuse lorsque le nombre de réplicats est faible.

Statistiques *t* modérées La statistique t modérée - parfois également appelée statistique B - d'un gène est un bien meilleur estimateur de la différence d'expression [194]. Elle est équivalente à une statistique t cependant qu'un facteur de

pénalité A est affecté au terme de variance.

$$t_g = \frac{\bar{y_g}}{A \times \sigma_g \sqrt{\frac{1}{n_g}}}$$

L'estimation de la valeur de A est le sujet de plusieurs travaux [195, 196, 180] dont les méthodes sont essentiellement fondés des réseaux bayésiens.

Ajustements Considérant m tests statistiques indépendants au risque α, la probabilité p de ne pas rejeter l'hypothèse nulle sur l'ensemble de ces tests est simplement le produit des probabilités individuelles $p = (1 - \alpha^m)$. Par exemple pour $\alpha = 0.5,\ m = 10$ alors $p = 0.6$. En d'autres termes, il existe 40% de risque qu'un test rejette l'hypothèse nulle par hasard alors qu'elle n'est que de 5% au niveau individuel. Ces résultats erronés sont appelés erreurs de type I.

Bonferroni Une première approche pour corriger les erreurs de type I est celle dite de Bonferroni [197]. Elle détermine le seuil de réjection des hypothèses π_i par un simple rapport $\pi_i < \alpha/m$. Cependant, dans le contexte des données d'expression par puce à ADN, m est trop important et la correction ainsi apportée est trop stringente.

Taux de Fausses Découvertes Le taux de fausses découvertes (*False Decovery Rate* - FDR) [198] est une méthode moins stringente que celle précédemment énoncée. Elle correspond à déterminer la probabilité d'hypothèses nulles injustement rejetées dans une liste d'hypothèses nulles rejetées.
Ceci signifie que dans une liste d'hypothèses nulles rejetées au risque α, une proportion α de ces hypothèses est rejetée au hasard.

3.5.2 Regroupements

Le fondement du regroupement des gènes, des échantillons ou des conditions expérimentales est de créer des ensembles homogènes de profiles d'expression. L'intérêt est multiple et varié tant il permet de définir des profiles types ou *classes* [199, 122], de diminuer la compléxité des données pour faciliter leur visualisation ou encore de détecter des valeurs extrêmes inclassables. Fondés sur la notion de dissimilarité, un grand nombre d'algorithmes de regroupements ont été apportés de la phylogénie aux études d'expression.

3.5.2.1 Distances et Dissimilarités

La représentation vectorielle dans l'*espace d'expression* de la matrice d'expression (cf page 79) permet l'usage de l'algèbre linéaire et par conséquent de calculer

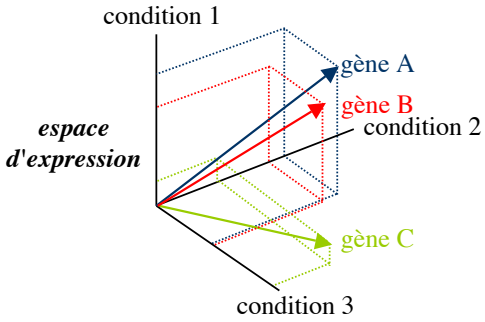


FIG. 3.14 – *Espace d'expression.* L'*espace d'expression* est l'espace dans lequel sont représentés les vecteurs d'expression. Chaque vecteur d'expression représente les valeurs d'un gène dans les différentes conditions expérimentales. Ainsi les gènes A et B ont une variation similaire de leur expression dans les trois conditions expérimentales alors que le gène C semble se comporter de façon différente.

la distance séparant deux vecteurs (gènes). Deux gènes dont les profiles d'expression sont similaires seront proches dans l'espace d'expression (cf. figure 3.14).
La distance de similarité des profiles d'expression est au centre des méthodes de regroupement. Nombre de méthodes ont été décrites afin de parvenir à une estimation de cette distance [200]. Les plus couramment utilisées dans le contexte des puces à ADN sont présentées.

Distances métriques Une distance métrique respecte les trois lois *sine qua none* suivantes :

- Elle doit être positive : $d_{i,j} \geq 0$;
- Elle doit être symétrique : $d_{i,j} = d_{j,i}$;
- Elle doit être triangulaire : $d_{i,k} \leq d_{i,j} + d_{j,k}$.

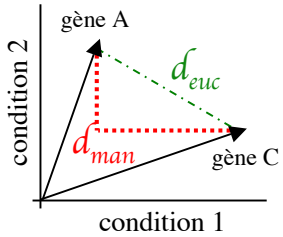


FIG. 3.15 – *Distances métriques.* La distance euclidienne entre les gènes A et B est représentée par les pointillés verts (d_{euc}) et la distance de Manhattan est représentée par pointillés rouges (d_{man}).

Distance Euclidienne La distance euclidienne est la plus usuelle des distances métriques. Elle correspond à la racine carré de la somme des carrés des différences des coordonnées dans chacune des N dimensions (cf. figure 3.15)

$$d^N_{euc}(A,C) = \sqrt{\sum_{i=1}^{N}(A_i - C_i)^2}$$

Manhattan La distance de Manhattan correspond à la somme des distances absolues entre les coordonnées des vecteurs d'expression dans chacune des N dimensions (cf. figure 3.15)

$$d^N_{man}(A,C) = \sum_{i=1}^{N}|A_i - C_i|$$

Bien que les distances métriques soient très intuitives, dans le contexte de l'expression des gènes leur utilisation est parfois injustifiée. C'est par exemple le cas dans les études comparatives dont l'intérêt porte davantage sur le changement des niveaux d'expression que sur les niveaux absolus de l'expression ; les distances non-métriques sont alors l'alternative (cf. figure 3.16).

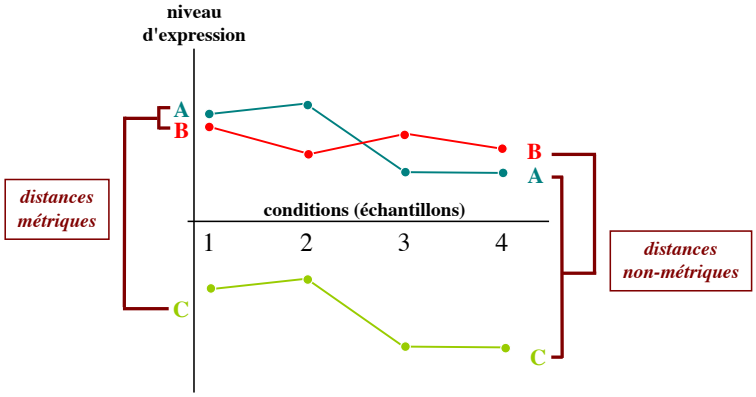


FIG. 3.16 – *Choix des distances.* Du choix de la distance utilisée dépendent les regroupements des gènes. Ainsi l'utilisation de distances métriques permettra de regrouper les gènes A et B dont les niveaux absolus d'expression sont proches, cependant que l'utilisation de distances non-métriques permettra de regrouper les gènes A et C dont les variations d'expressions sont identiques.

Distances non-métriques Les distances non-métriques sont des distances qui ne respectent pas au moins une loi des distances métriques.

Coefficients de corrélation linéaire de Pearson Fondé sur le calcul des distances angulaires entre les vecteurs, le coefficient de corrélation de PEARSON centré est en fait la covariance normalisées des coordonnées des vecteurs. En d'autres termes, deux vecteurs qui ont une variation d'expression similaire dans chacune des dimensions seront considérés en tant que voisins et seront alors très probablement affectés à un même ensemble par les procédures de regroupement (cf. figure 3.16).
Il est cependant parfois également intéressant de regrouper les vecteurs qui non seulement corrèlent mais également anti-corrèlent. Un facteur de transcription qui peut à la fois réprimer l'expression d'un gène mais activer celle d'un autre en est l'exemple. Dans ce cas l'utilisation du coefficient de corrélation de PEARSON au carré est une solution qui permettra de regrouper deux gènes certes anti-corrélés mais qui à l'évidence participent à un même évènement biologique.

Coefficients de corrélation des rangs de Spearman La distance entre deux vecteurs peut également être évaluée par le coefficient de corrélation des rangs de SPEARMAN. L'avantage de cette méthode de calcul est qu'elle est invariante aux changement monotones, *i.e.* elle ignore la magnitude des changements dès lors que les rangs sont conservés. Ainsi après avoir ordonné les valeurs de chacune des coordonnées des vecteurs dans chaque dimension, si l'ordre des rangs est le même pour chacun des vecteurs, le coefficient de corrélation est 1.

Gène	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5	Cond. 6
A	0	0	1	1	0	0
B	-1	-1	0	0	1	1
C	1	1	0	0	-1	-1

TAB. 3.1 – *Entropie d'information dans un espace discrétisé.* La notion d'entropie d'information est une mesure de l'incertitude de l'information obtenue à partir d'un vecteur d'expression afin de prédire le comportement des autres vecteurs. Cette matrice présentée ci-dessus comporte les vecteurs d'expression de trois gènes (A,B,C) dans six conditions. Les valeurs d'expression des gènes B et C se déterminent complètement l'une de l'autre. En effet lorsque B vaut -1 C vaut toujours 1, lorsque B vaut 0 C vaut toujours 0 et lorsque B vaut -1 C vaut toujours 1. Ainsi la connaissance des valeurs discrétisées de l'un réduit complétement l'incertitude des valeurs discrétisées de l'autre. Ces deux gènes sont donc extrêmement liés et leur entropie d'information est nulle.

Distances en espace discrétisé Il est parfois avantageux d'utiliser une matrice de données d'expression discrétisées [201, 202], pour laquelle les valeurs sont égales à $-1, 0, 1$ si le niveau d'expression corrrespondant décroît, est invariant ou croît respectivement.

Une mesure de cette distance est basée sur la notion d'information mutuelle ou *entropie d'information*. La notion d'entropie d'information est une mesure de l'incertitude de l'information obtenue à partir d'un vecteur d'expression afin de prédire le comportement des autres vecteurs. Ainsi si un second vecteur peut être complètement prédit à partir des coordonnées d'un premier vecteur, ces deux vecteurs seront alors très probablement affectés à un même ensemble par les procédures de regroupement (cf. table 3.1).

3.5.2.2 Algorithmes

Les algorithmes de regroupement sont organisés en plusieurs catégories non exclusives (cf. table 3.2). Parmi les très nombreux algorithmes de regroupements existants, seuls sont présentés les plus fréquemment utilisés en analyse d'expression par puces à ADN.

Algorithme	Supervision	Méthode	Représentation
ACP	non	agglomérative	plate
RHA	non	agglomérative	hiérarchique
RHD	non	divisive	hiérarchique
K-centroïdes	oui	divisive	plate
CAO	oui	divisive	plate

TAB. 3.2 – *Catégories des algorithmes de regroupement.* Les algorithmes de regroupement sont classés en fonctions de plusieurs critères. Le premier est la supervision de l'algorithme par des connaissances ante-analytique. Le K-centroïdes (*k-means*) et les cartes auto-organisées (CAO, *SOM*) par exemple nécessitent de définir avant l'analyse le nombre d'ensembles dans les quels les gènes seront placés. Ce nombre peut être estimé au regard des classes de tumeurs des échantillons par exemple mais également par une analyse statistique exploratoire (analyses factorielles) ou analytique par ré-échantillonage [203]. Le second concerne la méthode de création des ensembles. Si la méthode est divisive, l'algorithme considère l'ensemble des données comme un seul ensemble puis le divise jusqu'à l'obtention de groupes distincts. La méthode agglomérative est son antagoniste. Enfin le troisième critère est la représentation. Elle peut être hiérarchisée ou plate. Dans le premier cas comme pour les regroupements hiérarchisés agglomératifs (RHA) ou divisifs (RHD), les ensembles sont hiérarchisés les uns par rapport aux autres. Le diagramme de *Venn* est un exemple de représentation non hiérarchisée des ensembles.

Analyses factorielles

Analyse en composante principale L'analyse en composante principale (ACP, *CPA*) est une méthode des statistiques exploratives dont l'objet est de

diminuer le nombre des dimensions de l'espace d'expression sans perte d'information significative.
La réduction des dimensions est fondée sur la détermination de nouvelles directions ou *vecteurs-propre* dans l'espace d'expression (cf. figure 3.17). Les termes *gènes-propre* ou *conditions-propre* sont parfois employés dans le contexte des études d'expression génique.
Les vecteurs-propre et la valeur-propre associée sont calculés sur la matrice de covariances des données et représentent en conséquence la variabilités des données. Ainsi le vecteur-propre possédant la valeur-propre maximale est le vecteur qui représente le mieux la variabilité des données. Les valeurs dont la variabilité est minimale sont négligées au profit de celles dont la variabilité est maximale. Il est fréquent que quasi-totalité de la variabilité des données soit représentée sur trois directions, ce qui permet alors de projeter les données dans un espace à trois dimensions.
Souvent utilisée dans les études de données d'expression [204, 205] et proposée

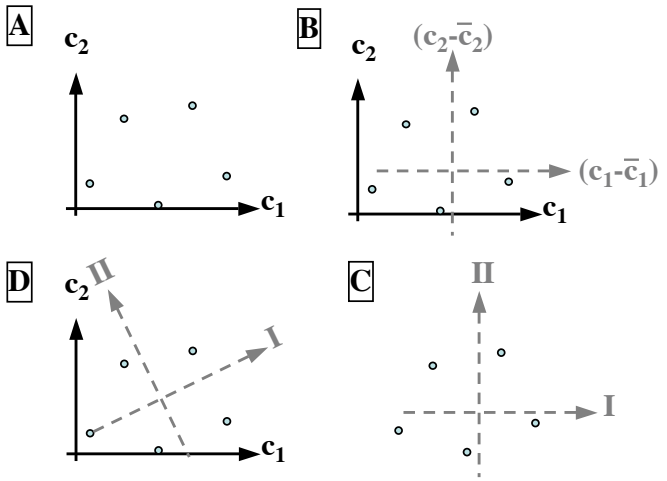


FIG. 3.17 – *Analyse des Composante Principale.*(A) Cinq gènes sont projetés dans l'espace d'expression d'axes c_1 et c_2 représentants les conditions expérimentales (échantillons). (B) Après recentrement des coordonnées des gènes, ces derniers sont maintenant projetés dans l'espace d'expression dont les axes sont $(c_1 - \overline{c_1})$ et $(c_2 - \overline{c_2})$ tracés en pointillés. (C) Les gènes sont projetés sur les *vecteurs-propre* I et II calculés sur la matrice de covariance des données. (D) Les deux systèmes d'axes peuvent alors être superposés par rotation du second.

dans la quasi-totalité des logiciels d'analyse, il important de relevé certaines caractéristiques notables concernant l'ACP.
Les vecteurs-propre sont orthogonaux et cela a au moins deux conséquences sur les résultats de l'analyse : d'abord une inefficacité dans certaines compositions de données. En effet la transformation effectuée lors du changement du système

de coordonnées par rapport au système initial est une rotation des axes d'origines, eux-même orthogonaux. Or il arrive pour certaines données que cette solution ne soit pas optimale [206] (cf. figure 3.18).Deux gènes proches de deux vecteur-propre différents ont une corrélation nulle et deux gènes projetés aux deux extrémités d'un même vecteur-propre sont anti-corrélés. Ces interprétations sont par ailleurs assujetties à une pondération tenant compte de la valeur-propre associée au vecteur-propre.
Les directions des nouveaux axes sont déterminées sur la variance des données et ne tiennent par exemple pas compte de leur classe. Ainsi des données de même variance mais de classe différentes auront les mêmes coordonnées sur les vecteurs-propre.

Analyse factorielle des correspondances À l'instar de l'ACP, l'analyse factorielle des correspondances (AFC) est une technique des statistiques descriptives dont l'objectif est également de diminuer le nombre de dimensions des données pour faciliter leur interprétation. Cependant lorsque l'ACP permet uniquement la projection exclusive des variables (gènes ou échantillons) dans l'espace définit par les vecteurs-propre, l'AFC permet la représentation simultanée des variables révélant ainsi les associations intra- et inter-variables [207, 208].

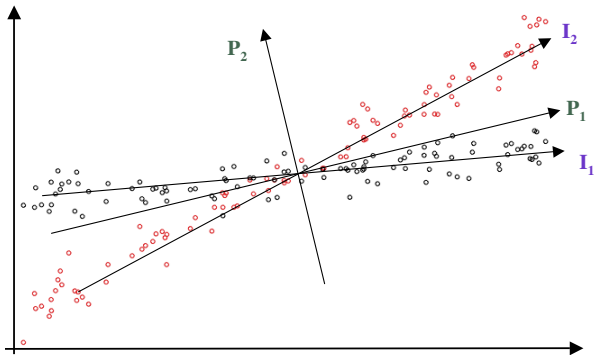


FIG. 3.18 – *Analyse en Composante Indépendante.* Les facteurs (I_1, I_2) déterminés par l'ICA sont une meilleure représentation des données que celle proposée par les facteurs (P_1, P_2) de l'ACP tenus à l'orthogonalité.

Analyse des composantes indépendantes L'analyse en composante indépendante (*ICA*) est une technique capable de considérer des dépendances statistiques telles que l'inclinaison des projections et la forme de la distribution (cf. figure 3.18) [205, 209, 210] parfois mieux adaptée que l'ACP ou l'AFP fondées uniquement sur la variance des données et dont les directions sont toujours orthogonales.

Regroupements hiérarchiques Issus des études de taxinomie [211], les regroupements hiérarchiques sont fréquemment utilisés dans les études d'expression génique. Cette méthode de regroupement propose une organisation hiérarchisée des ensembles de gènes souvent représentés par un arbre (cf. figure 3.19 B).
Dans sa forme agglomérative, l'algorithme est initié par la composition d'une matrice des distances à partir des vecteurs d'expression. Chaque vecteur forme un ensemble singleton. Les deux ensembles les plus proches sont alors assemblé pour former un nouel ensemble. Les nouvelles distances entre cet ensemble et les autres alors est alors de nouveau calculée. Cette étape est répétée jusqu'a l'obtention d'un ensemble comprenant la totalité des vecteurs.
Les désavantages de ces méthodes sont multiples. Le premier est la faible qualité de la représentation des éléments. En effet plus la taille de l'ensemble augmente moins il représente les éléments qui le constituent. Le second est l'impossibilité de correction des erreurs d'assignation. Enfin le dernier est la multiplicité des représentations. Ainsi les dendogrammes qui représentent les ensembles déterminés ne sont pas uniques et une même discrimination peut être représentée par plusieurs arbres [212].

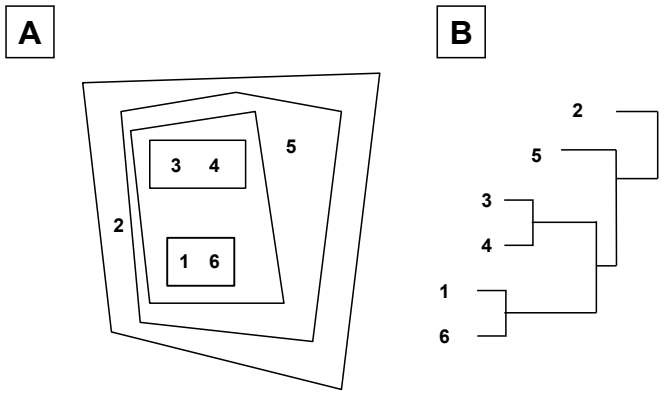


FIG. 3.19 – *Regroupement hiérarchique.* Le regroupement hiérachique des données peut être présenté par un système d'ensembles et de sous-ensembles (A) ou par un arbre dont la longueur des branches est fonction de la distance d'une feuille à une autre (B).

Agglomératifs [204] ou divisifs [213, 199] ces algorithmes s'organisent en au moins cinq méthodes de détermination des ensembles.

Liaison simple La frontière entre deux ensembles est fondée sur la distance entre les deux vecteurs les plus proches de chacun des ensembles (cf. figure 3.19(1).

Liaison complète La frontière entre deux ensembles est fondée sur la distance entre les deux vecteurs les plus éloignés de chacun des ensembles (cf. figure 3.19(1)).

Liaison moyenne La frontière entre deux ensembles est fondée sur une distance moyenne (*unweighted paired-group method*,UMPGA) correspondant à la moyenne de la totalité des vecteurs de chacun des ensembles (cf. figure 3.19(2)). Une variante consiste à remplacer la moyenne par la valeur médiane.

Centroïdes La frontière entre deux ensembles est fondée sur la distance entre les centres de gravité de chacun des ensembles (cf. figure 3.19(3)).

Moyenne pondérée Il s'agit d'une variation pondérée de l'UMPGA. Cette méthode tient compte de la taille des ensembles, *i.e.* le nombre de vecteurs.

Ward Déterminer si un vecteur appartient à un ensemble par la méthode de Ward [214] consiste à calculer un score pour chaque ensemble fondé sur les écart-types des ensembles et donc représentatif de la variabilité de l'ensemble. Un vecteur est autorisé à intégrer un ensemble s'il est celui qui augmente le moins le score de cet ensemble.

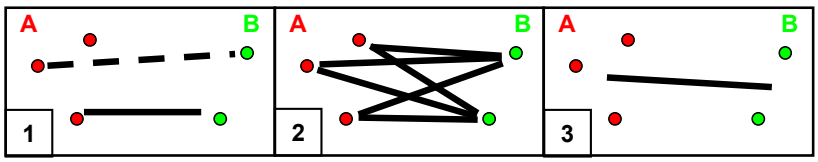


Fig. 3.20 – *Algorithmes de regroupement hiérarchique.* Différents algorithmes de regroupement permettent de définir la frontière entre les ensembles A et B. (1) Le *single linkage* est fondé sur la distance entre les vecteurs les plus proches alors que le *complete linkage* est fondé sur la distance netre les vecteurs les plus éloignés. (2) La frontière entre les deux ensembles est fondée sur une distance moyenne (*unweighted paired-group method*,UMPGA) correspondant à la moyenne de la totalité des vecteurs de chacun des ensembles. (3) La frontière entre deux ensembles est fondée sur la distance entre les centres de gravité (centroïdes) de chacun des ensembles.

Regroupement par partitions

K-means Le *k-means* un algorithme de partitionnement des données. La détermination des partitions constitue la phase d'initiation de l'algorithme et consiste en l'affectation aléatoire d'un nombre déterminé *a priori* de centres. Initialement décrits dans les premières versions de l'algorithme comme les centroïdes des partitions [215], ces centres sont parfois des médoïdes, *i.e.* le meilleur objet représentatif de chaque partition [216].

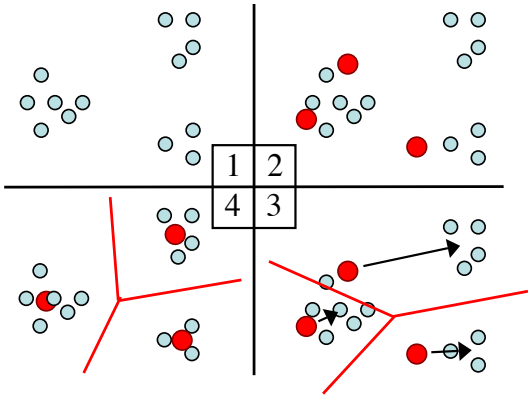


FIG. 3.21 – *Regroupement par K-means.* (1) Les vecteurs sont ici présentés dans une espace à deux dimensions. (2) La phase d'initiation de l'algorithme consiste en l'affectation aléatoire d'un nombre déterminé *a priori* de centres (points rouges). (3) Cette phase est suivie par une phase d'affectation et de détermination des partitions. Chaque objet voit alors sa distance aux centres calculée afin de l'intégrer à la partition dont il est le plus proche. Les partitions sont ainsi une première fois définies (lignes rouges). (4) Cependant les nouveaux centres des nouvelles partitions sont calculés itérativement et les partitions sont de nouveau définies. Lorsque les centres ne se déplacent plus de manière significative les partitions sont stabilisées.

Après la phase d'initiation chaque objet voit alors sa distance aux centres calculée afin de l'intégrer à la partition dont il est le plus proche. Les nouveaux centres des nouvelles partitions sont calculés itérativement jusqu'à ce qu'ils ne se déplacent plus de manière significative (cf. figure 3.21).
Bien que simple d'utilisation et très fréquemment utilisé, le partitionnement par *k-means* possède quelques points faibles conséquences particulières de la phase d'initiation et de la détermination aléatoire des centres. En effet si les centres sont dans une zone peu peuplée, ils conduiront à la formation de partitions vides. Par ailleurs, un même jeu de données traité par *k-means* conduit fréquemment à plusieurs partionnements différents. Une solution à ce dernier problème est l'utilisation d'outils de validation statistiques tel que le *bootstrap* [217] qui donne un indice de confiance aux partition formées. Cette technique est basée sur la

modélisatin des données observées. De ce modèle sont générés des jeux de données aléatoires, lesquels subissent chacun un partitionnement. Les partitions qui varient peu entre les données observées et les données générées sont de confiance, les autres non.

Cartes auto-organisées, (*Self Organised Maps*) Algorithme de la famille des réseaux de neurones [218] et adapté à l'analyse des donnnées d'expression [219, 220], les cartes auto-organisées sont également fondées sur un partitionnement des données. Deux notions directrices conduisent le partitionnement des données. La notion de *voisinage* et la notion *gagneur absolu*.
Dans la phase initiale la géométrie des partitions est pré-définie et un vecteur référence est généré aléatoirement pour chacune de ces partitions. Chacun des vecteurs référence est lié à l'enemble des données.
Durant la phase d'apprentissage, l'ensemble des données est présenté à l'ensemble des vecteurs références dans un ordre aléatoire et plusieurs fois. Durant chacune des ces itérations (plusieurs milliers), le vecteur référence le plus proche de la données présentée est le gagnant absolu et est recalculé pour tenir compte de sa similarité avec la donnée. Par ailleurs, la modification du vecteur référence influence les vecteurs références voisins. La phase d'apprentissage est terminée lorsque la présentation des données ne modifie plus la géométrie des partitions.
La dernière phase affecte les données à chacune des partitions dont elles sont les plus proches.

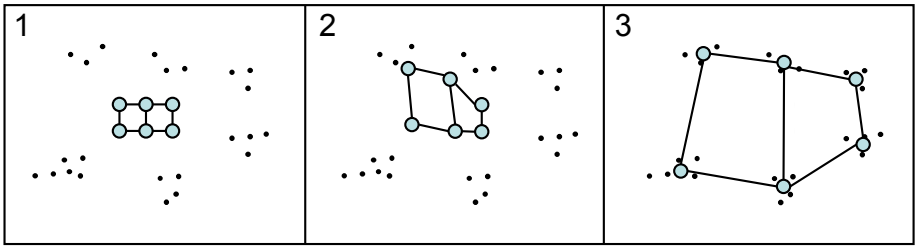


FIG. 3.22 – *Cartes auto-organisées.* (1) Dans la phase initiale la géométrie des partitions est pré-définie et un vecteur référence est généré aléatoirement pour chacune de ces partitions. Chacun des vecteurs référence est lié à l'ensemble des données. (2) Durant la phase d'apprentissage, l'ensemble des données est présentée à l'ensemble des vecteurs références dans un ordre aléatoire et plusieurs fois. Durant chacune des ces itérations (plusieurs milliers), le vecteur référence le plus proche de la données présentée est le gagnant absolu et est recalculé pour tenir compte de sa similarité avec la donnée. Par ailleurs, la modification du vecteur référence influence les vecteur référence voisins. (3) La phase d'apprentissage est terminée lorsque la présentation des données ne modifie plus la géométrie des partitions.

Partitions floues Le principal désavantage des deux méthodes déterministes précédentes est le partage strict des données entre les différentes partitions. Il est en effet possible que des gènes appartiennent en réalité à plusieurs partitions. Une solution est d'utiliser des partitions flouees (*fuzzy clustering*). L'idée consiste à affecter aux données des indices ou probabilités bayesiennes en fonction de leur affinité avec chacune des partitions. Ainsi les données n'appartiennent pas exclusivement à une partition mais appartiennent plus probablement à certaines partitions qu'à d'autres [221].

3.5.3 Discrimination

Alors que les groupes issus des méthodes de regroupement précédentes supportent difficilement l'addition de nouvelles données sans subir une modification de leur structure, les méthodes de discrimination, au contraire se fondent sur la formation de groupes modèles ou discriminateurs permettant la discrimination de nouvelles données [222].

3.5.3.1 Définition des classes

Considérant la matrice X (cf. page 78) d'expression des gènes, la matrice des classes est définie comme suit

$$\begin{pmatrix} L_1 & L_2 & L_3 & \cdots & L_m \\ x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{pmatrix}$$

Notons chacune des colonnes de cette matrice

$$A_1, ..., A_m \; avec \; A_1 = (x_{11}, ..., x_{n1}), ..., A_m = (x_{1m}, ..., x_{nm})$$

Ces colonnes sont chacune affectée d'une étiquette ou *variable de classe* $L_i \in \{1, -1\}$ indiquant la classe d'appartenance. Le vecteur $(L_1, ..., L_m)$ est le *vecteur de classes*.
Un algorithme de discrimination a donc la propriété de prédire la valeur de L_i connaissant A_i. Pour cela l'algorithme est composé d'une phase d'apprentissage qui lui premet d'optimiser les paramètres de prédiction. Cet apprentissage est supervisé, *i.e.* qu'il est effectué grâce à des données *a priori* classées sur critères biologiques ou anatomopathologiques (échantillons,types de tumeurs...). Cette phase est suivie par une phase de test utilisant des méthodes de ré-echantillonnage ou d'aggrégation des clusters (*bagging*) [223] afin de vérifier la justesse des prédictions.

Cette étape franchie, l'algorithme est prêt à prédire la variable de classe L_{m+1} correspondant au profile A_{m+1}.

3.5.3.2 Discrimination linéaire

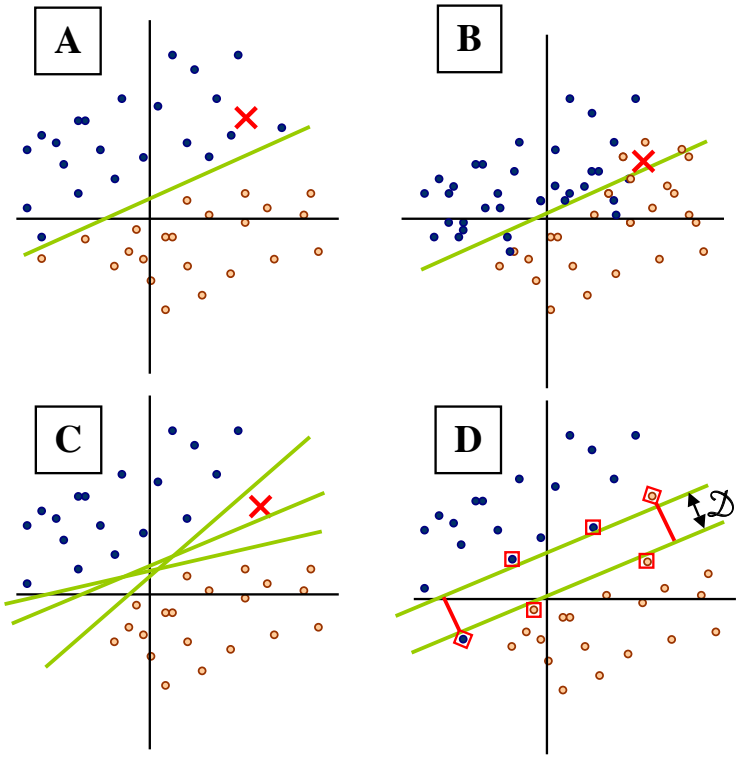


FIG. 3.23 – *Discriminants linéaires.* Les points repésentent les profiles d'expression dans un espace à deux dimensions. Les points marrons appartiennent à la classe 1 et les bleus à la classe 2. (A) La droite permet une séparations des données et de classer correctement e nouveau profile d'expression représenyté par une croix rouge. (B) Les données sont inséparables dans un espace à deux dimensions et classé avec assurance le nouveau profil est impossible. (C) Plusieurs droites permettent de définir une séparation entre les profiles. En conséquence il est impossible de choisir la classe à laquelle appartient le nouveau profil. (D) Afin de séparer les profiles, l'espace discriminant D de plus large entre les points est déterminé en limitant la somme des distances aux point placés dans la mauvaise classe (carrés rouges). Ces points sont appelés *vecteurs supports*

.

Dans un espace à n dimensions, si $n = 2$, la discrimination linéaire partage les données en classes par la détermination d'une droite séparatrice (cf. figure 3.23 A). Déterminer l'équation de cette droite est aisé et en conséquence la discrimination

des données l'est tout autant. Une solution simple et fréquente pour déterminer cette droite est l'utilisation de régressions linéaires par moindre carrés. Le principe consiste à rechercher la droite pour laquelle la distance à chacune des données est minimale.
Cette approche est étendue dans les espaces à plus de deux dimensions en déterminant les paramètres d'un plan si $n = 3$ ou d'un *hyperplan* si $n > 3$.
Il est pour autant des cas de figures pour lesquels une telle méthode est incapable de discriminer les données ; lorsque par exemple une droite ne parvient pas à séparer les données ou au contraire lorsque plusieurs droites ont la propriété de séparer les données (cf. figure 3.23 B et C).

3.5.3.3 K plus proches voisins

Une alternative simple à la discrimination linéaire est la discrimination par les *K plus proches voisins*. Ainsi l'algorithme désigne la classe d'affectation d'un nouveau profil en le comparant aux K profiles les plus proches (déterminés par calcul de la distance Euclidienne par exemple). La classe désignée est alors la classe majoritaire parmis celles des K plus proches voisins [121].

3.5.3.4 Séparateur à Vaste Marges

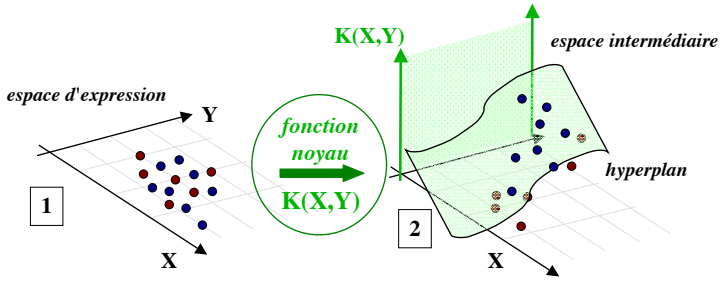


FIG. 3.24 – *Séparateur à Vaste Marges.* (1) La discrimination des données présentées ici dans l'espace d'expression est impossible. (2) La solution consiste à les transposer dans un espace de dimension supérieure afin de déterminer un hyperplan capable de discriminer efficacement ces données. La transformation est définie par une fonction noyau K.

Le principe de base des Séparateurs à Vaste Marge (SVM, *Support vector Machine*) [224], revient à ramener le problème de la discrimination linéaire à la recherche d'un hyperplan optimal. Deux idées permettent d'atteindre cet objectif :

- La première consiste à définir et optimiser les paramètres de l'hyperplan séparateur contrôlés par les vecteurs supports ;
- La seconde consiste à définir une *fonction noyau* (*kernel*) introduite dans le produit scalaire des vecteurs. Cette fonction est le passage vers un *espace*

intermédiaire qui permet la résolution de problèmes linéaire dans un espace non linéaire de plus grande dimension. La fonction $K(X, Y) = (X.Y + 1)$ est un exemple fréquemment utilisé de fonction noyau (cf. figure 3.24).

Cet algorithme très performant surpasse fréquemment les autres méthodes de discrimination. Il est pour cette raison souvent utilisé dans les études d'expression génique tant pour classer les gènes [225] que pour classer les échantillons [226].

3.5.3.5 Arbres décisionnels

La construction d'arbres binaires a pour objectif de modéliser la discrimination de données quantitatives grâce à des données qualitatives [227]. Visuellement simple à interpréter, ce sont des outils efficaces pour l'aide à la decision, d'où leur nom d'*arbres décisionnels*.

Le principe consiste schématiquement en une série de questions fermées (réponses binaires) permettant de diviser les données. Chacune de ces questions est définie en fonction de la réponse à la question précédente définissant ainsi une séquence de noeuds. À chaque noeud correspond une dichotomie de la variable. La croissance de l'arbre se termine lorqu'il devient homogène *i.e.* lorsqu'il n'existe plus de division admissible.

Cette méthode est particulièrement adaptée aux données de grande taille et son utilisation fût naturellement proposée pour les études d'expression génique [228].

Le grand nombre de méthodes pour l'analyse des données d'expression issues des puces à ADN est à la mesure des efforts développés pour proposer de nouveaux outils, notamment dans le domaine des licences gratuites avec la suite logicielle telle BioConductor [229]. Ces outils associés à une étude statistique pré-analytique (schéma expérimental) et post-analytique (transformation, regroupement et discrimination) permettent d'exploiter le complet potentiel des puces à ADN.

Seconde Partie

Chapitre 4

Le transcriptome hépatique de la phase aiguë in vivo

L'analyse du transcriptome nous interesse à trois titres :
- Le premier vient du fait que le profile d'expression des gènes et donc le profile d'abondance des ARNm est utilisé dans le contexte pathologique comme une signature de la pathologie ;
- Le second vient du fait que les modifications d'abondance que subissent les ARNm ont parfois pour conséquence des modifications sur les protéines directement inpliquées dans la pathologie ;
- Le dernier intérêt vient enfin du fait que l'étude à grande échelle des modifications au niveau des ARNm permet d'intégrer l'ensemble des protagonistes impliqués dans la pathologie.

Le foie contient un grand nombre de gènes transcripts. Ces transcripts permettent la production de protéines qui participent à des évènements vitaux pour l'organisme mais également à des processus spécifiques du foie telle la régénération hépatique.
En réponse à un abus d'alcool chronique ou à une infection virale hépatique, le foie subit alors des modifications tissulaires importantes dont les conséquences sont la cirrhose et le carcinome hépatocellulaire.
Par ailleurs la phase aiguë de l'inflammation modifie positivement ou négativement l'expression hépatique des gènes de l'immunité innée qui codent par exemple pour les protéines plasmatiques de la phase aiguë.

Le foie est donc un organe majeur et l'étude de son transcriptome peut aider à définir des nouveaux marqueurs, de nouvelles voies de régulations ainsi que des nouvelles fonctions protéiques.

Puisqu'une vue globale du transcriptome hépatique humain *in vivo* au cours de l'inflammation systémique aiguë n'a encore jamais été proposée, nous avons développé *Liverpool*, une puce à ADN représentant approximativement 10000 gènes et recouvrant ainsi le transcriptome hépatique.

Associé à *Liverpool*, jài développé *LiverTools* qui est un ensemble d'outls composé d'un réseau de serveurs et de clients, d'une base de données pour emmagasinner l'ensemble des informations inhérentes aux expériences menées grâce à *Liverpool* et de logiciels d'analyse statistique.

Altered Gene Expression in Acute Systemic Inflammation Detected by Complete Coverage of the Human Liver Transcriptome

Cédric Coulouarn,[1] Grégory Lefebvre,[1] Céline Derambure,[1] Thierry Lequerre,[1,2] Michel Scotte,[1,3] Arnaud Francois,[4] Dominique Cellier,[5] Maryvonne Daveau,[1] and Jean-Philippe Salier[1]

The goal of the current study was to provide complete coverage of the liver transcriptome with human probes corresponding to every gene expressed in embryonic, adult, and/or cancerous liver. We developed dedicated tools, namely, the *Liverpool* nylon array of complementary DNA (cDNA) probes for approximately 10,000 nonredundant genes and the *LiverTools* database. Inflammation-induced transcriptome changes were studied in liver tissue samples from patients with an acute systemic inflammation and from control subjects. One hundred and fifty-four messenger RNAs (mRNA) correlated statistically with the extent of inflammation. Of these, 134 mRNA samples were not associated previously with an acute-phase (AP) response. The hepatocyte origin and proinflammatory cytokine responsiveness of these mRNAs were confirmed by quantitative reverse-transcription polymerase chain reaction (Q-RT-PCR) in cytokine-challenged hepatoma cells. The corresponding gene promoters were enriched in potential binding sites for inflammation-driven transcription factors in the liver. Some of the corresponding proteins may provide novel blood markers of clinical relevance. The mRNAs whose level is most correlated with the AP extent ($P < .05$) were enriched in intracellular signaling molecules, transcription factors, glycosylation enzymes, and up-regulated plasma proteins. In conclusion, the hepatocyte responded to the AP extent by fine tuning some mRNA levels, controlling most, if not all, intracellular events from early signaling to the final secretion of proteins involved in innate immunity. *Supplementary material for this article can be found on the HEPATOLOGY website (http://interscience.wiley.com/jpages/0270-9139/suppmat/index.html).* (HEPATOLOGY 2004;39:353–364.)

Three major objectives drive the current interest in human transcriptome analysis. First, messenger RNA (mRNA) profiling in a pathologic context is used as a molecular signature of diagnostic and/or prognostic relevance, regardless of the functions of the cognate proteins. Second, some mRNA changes may point to proteins that are directly involved in a disease process. Third, mRNA changes can be studied in an integrative context in which the so-called "guilt-by-association" approach aims at identification of all the participants in a given signaling cascade or metabolic pathway.[1–3] Progress in any of these areas with array technology[4] depends on probe diversity. Therefore, the use of a pan-genomic set of probes is the ideal option. However, in practice, such an array is impractical given the current uncertainties about the number and identity of human genes,[5,6] the exponentially growing complexity of data analysis that results from a linearly increasing number of probes,[7] and the cost of a pan-genomic probe set. Nevertheless, maintaining a high probe diversity for studies in a given cell type or tissue context is required for an integrative approach and should also help discover many candidate genes whose hallmark

Abbreviations: mRNA, messenger RNA; AP, acute phase; APP, acute phase protein; cDNA, complementary DNA; Q-RT-PCR, quantitative reverse-transcription-polymerase chain reaction; Hs., Homo sapiens; TF, transcription factor; CRP, C-reactive protein; SAA, serum amyloid A; Clqβ, β chain of complement Clq; TSF, transferrin; MIAME, minimum information about a microarray experiment; ALB, albumin; IL, interleukin; C/EBP, CCAAT-enhancer binding protein; NK4, natural killer cell transcript 4; IGFBP2, insulin-like growth factor binding protein 2; MAP4K4, mitogen-activated protein kinase kinase kinase kinase 4; NFκB, nuclear factor κB; STAT, signal transducer and activator of transcription; TRAF5, tumor necrosis factor receptor-associated factor 5.

From [1]INSERM Unité 519 and Faculté de Médecine-Pharmacie, Institut Fédératif de Recherches Multidisciplinaires sur les Peptides, Rouen, France; [2]Service de Rhumatologie, [3]Service de Chirurgie Générale et Digestive, and [4]Département de Pathologie, Centre Hospitalier Universitaire, Rouen, France; and [5]Laboratoire de Mathématiques Raphaël Salem and UMR CNRS 6085, Université de Rouen, Rouen, France.

Received July 8, 2003; accepted November 2, 2003.

C.C. and G.L. are the recipients of a fellowship from the French Ministry for Research and C.C. and C.D. are the recipients of a fellowship from Association de Recherche sur le Cancer and Ligue contre le Cancer, respectively. This work was supported, in part, by grants from Association de Recherche sur le Cancer and Groupement des Entreprises Françaises dans la LUtte contre le Cancer (GEFLUC) to J.-P.S.

Address reprint requests to: Jean-Philippe Salier, INSERM Unité 519, Faculté de Médecine-Pharmacie, 22 Bvd. Gambetta, 76183 Rouen cedex, France. E-mail Jean-Philippe.Salier@univ-rouen.fr; fax: +33-235-14-85-41.

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI 10.1002/hep.20052

appears to be a tissue-restricted expression.[8] Overall, a probe selection that results in a virtually complete coverage of the transcriptome in a given cell type or tissue is likely to be the best choice for the time being. Yet, such a tool has seldom been developed.

The liver contains a large number of transcribed genes whose products participate in a vast array of vital and organ-specific functions as well as organ-restricted properties such as a high capacity to regenerate.[9] In response to chronic alcohol abuse or hepatitis B or C virus infection, the liver may undergo major tissue modifications that result in cirrhosis and subsequent hepatocellular carcinoma.[10] Therefore, the liver ranks high among those whose transcriptome richness may help decipher as yet unknown disease markers, critical gene regulations, and novel protein functions.[11,12] In addition, the acute phase (AP) of a systemic inflammation up-regulates or down-regulates many liver-expressed genes involved in innate immunity and coding, for instance, for the positive or negative plasma acute phase proteins (APP).[13–16] However, a global view of the AP-induced changes in the liver transcriptome has not yet been obtained in humans *in vivo*. We report the development of an array based on selected human complementary DNA (cDNA) probes that correspond to approximately 10,000 nonredundant genes and specifically cover the liver transcriptome. This allowed us to identify the liver mRNAs whose abundance best correlates with the extent of an acute, systemic inflammation in humans.

Materials and Methods

Human Subjects and RNA Sources. Total RNA samples from human fetal liver specimens (15-24 week-old fetuses) or adult brain were obtained from Clontech (Palo Alto, CA). Sections of normal tissue taken from surgically resected livers were obtained from the digestive surgery unit of Charles Nicolle Hospital (Rouen, France) under strict anonymity. The diagnosis was either a primary tumor or a hepatic metastasis that was detected in the follow-up of a nonhepatic carcinoma. A matched blood sample was obtained before surgery. In some instances, a hepatocyte-enriched fraction (<3% contamination by nonparenchymal cells) was immediately prepared as described previously.[17] Other tissue samples were obtained from several units in the same hospital. According to French law and ethical guidelines, no informed consent is requested before analysis of RNA samples from resected tissue specimens that would otherwise be discarded. The culture and stimulation of Hep3B hepatoma cells and quantitative reverse- transcription polymerase chain reaction (Q-RT-PCR) of mRNA are fully detailed as a supplementary material on the HEPATOLOGY website (http://interscience.wiley.com/jpages/0270-9139/suppmat/index.html), as well as on our website (www.lille.inserm.fr/u519/coulouarnetal03.html).

Selection of Homo sapiens Clusters and Promoter Sequences. From the Unigene database (ftp://ftp.ncbi.nih.gov/repository/Unigene), a parsing of *Homo sapiens* (Hs.) data files (build 129, June 2003) was performed with locally imported flat files. The parsing algorithm implemented in PERL (available upon request) allowed us to select a series of Hs. clusters with the single following criterion: expression must occur at least in the liver. Elsewhere, the Unigene Library Browser along with the whole set of cDNA libraries listed in Unigene allowed us to select every library that obeyed the following single criterion: the library must be constructed from a human liver-related tissue sample. Every Hs. cluster contained in any such library was first retained. A PERL program was used to establish a list of nonredundant Hs. clusters from all the above sources. Finally, a bibliographic search made with standard online tools still identified further Hs. clusters. Our final Hs. cluster selection also encompassed the cDNAs for 38 housekeeping genes.[18] It is available upon request.

Promoter sequences from Hs. cluster-defined genes were retrieved from mRNA / gene alignments with the Evidence Viewer tool (http://www.ncbi.nlm.nih.gov/LocusLink/). A set of control promoter sequences were made of Hs. cluster-defined genes that were not AP responsive in the current study and were chosen on the sole basis of promoter sequence availability. This analysis was performed over the first 5 kb of DNA sequence upstream of the transcription start site, provided these 5-kb sequence were available. Potential transcription factor (TF) binding sites were searched in the promoter sequences with the MatInspector and TRANSFAC tools.[19]

Probe Selection, Array Preparation, and Hybridization. The selection, amplification, and arraying of cDNA clones, [α^{33}P]dCTP labeling, and hybridization of total RNAs and image analysis are detailed as a supplement on the HEPATOLOGY website and our website.

Data Normalization, Filtering, Statistical Analysis, and Final Data Handling. To allow for comparison between images, normalization was based on the mean of the signals provided by the complete set of spots per image. All data in the current study were obtained from at least three separate hybridizations per RNA sample and the genes were identified as expressed if at least two hybridizations provided a positive signal. For every probe, the signals obtained under two different conditions (*i.e.*, A vs. B) were expressed as the difference (normalized signal in condition A − normalized signal in condition B)

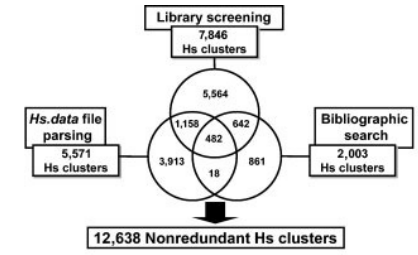
and considered to be significantly induced or repressed (folds) if this difference was outside a CI ($P < .05$) calculated[20] from the entire data set. To select the liver mRNAs that were regulated in patients with an acute inflammation versus controls, the value for any given mRNA from every individual with an inflammation was compared with the mean value obtained from the control set. All statistical analyses were performed with the R software.[21] Hierarchical clustering was performed with the Cluster and Tree View software and the uncentered correlation and complete linkage clustering options were used.[22]

LiverTools Database. This database utilizes a MySQL relational database server, an Apache web server, and PHP. Accordingly, it can be queried *via* an internet browser under any operating system. The data are gathered within constraint-linked tables. A module written in the PERL language allows the set of data contained in the "cDNA probes / array design" section of *LiverTools* to be updated weekly by a connection to NCBI (National Center for Biotechnology Information). *LiverTools* complies with minimum information about a microarray experiment (MIAME) recommendations[23] and is accessible upon request. The functions of proteins encoded by various mRNAs (see the Results section) were retrieved from the LocusLink (www.ncbi.nlm.nih.gov/LocusLink) and OMIM (www.ncbi.nlm.nih.gov/omim/) databases.

Results

Liverpool and LiverTools: a Liver-Oriented cDNA Array and Associated Database. Our goal was to provide extensive coverage of the human liver transcriptome. We used a single stringent criterion to select as many nonredundant genes as possible, namely, they must be expressed at least in the human liver under normal or pathologic conditions. As shown in Fig. 1, the results of our *in silico* searches (see Materials and Methods) eventually provided 12,638 nonredundant Unigene Hs. clusters. The corresponding genes evenly cover all human chromosomes as their locations, when known, strongly correlate ($r = 0.96$, n = 5,765, $P < 10^{-4}$) with the overall gene frequency per human chromosome (www.ncbi.nlm.nih.gov/genome/guide/human/HsStats.html) and do not indicate any preference for given chromosomes (not detailed). This is also true (data not shown) for a subset of 805 located Hs. clusters with a liver-restricted expression (detailed below), which is in keeping with the lack of tissue-specific gene clustering on chromosomes.[24]

cDNA clones covering the Hs. clusters were purchased from the IMAGE consortium (RZPD, Berlin, Germany). However, for a limited number of Hs. clusters, no IMAGE clone was available and 15.2% of the 12,493 cDNA clones that were purchased did not provide a satisfactory PCR product. Eventually, our nylon array, dubbed *Liverpool*, typically harbored 13,824 probes. They included a set of usable cDNA probes covering 9,858 nonredundant Hs. clusters (Unigene build 155) as well as a large set of control spots (Fig. 1). Quality controls were performed. First, because many IMAGE clones are suspected of misidentification, a limited number (n = 686) of human cDNA clones was controlled by end sequencing and an overall misidentification rate was calculated. This rate was 6.9% (out of 131 resequenced clones) for the subset of clones pertaining to a limited IMAGE population previously sequence verified at NCBI (ftp://image.llnl.gov/image/clones_verified) whereas it was 11.9% (out of 555 resequenced clones) for the subset of clones pertaining to the major set of non-NCBI–verified clones. These rates compare quite favorably with earlier estimates made from other IMAGE clones.[4] Second, our option of a global background substraction is straightforward but requires a reproducible signal to be obtained with multiple copies of



Liverpool		
cDNA clones		**Number of spots**
Associated to an Hs number		**11,604**
	Nonredundant clusters	9,858
	Expressed Sequence Tag	3,650
	With a chromosomal location	9,083
	With a function known or inferred	3,983
Non-associated to an Hs number		**889**
Controls		
Spike	8 *A. thaliana* cDNAs	**44**
Negative	Blank; insert-free plasmid; matrix-free PCR; genomic DNA; poly(A) tract	**1,287**

Fig. 1. *Liverpool*, a liver-oriented cDNA array. (**Upper panel**) Selection of human genes with hepatic expression. A Venn diagram shows the overlaps among three sets of genes with an expression in the liver at least, as judged from information included in Hs. data files (Unigene build 129), *in silico* screening of hepatic cDNA libraries (the resulting numbers of nonredundant Hs. clusters: 7,402 in 23 human adult liver libraries; 490 in three infant liver libraries; 1,340 in five fetal liver libraries; 4,082 in 12 hepatocellular carcinoma libraries; 1,340 in two HepG2 libraries), and bibliographic analyses. A final series of 12,638 nonredundant Hs. clusters was selected. (**Lower panel**) A list and properties of representative human cDNA clone(s) for every selected Hs. cluster and control clones arrayed over every *Liverpool* filter.

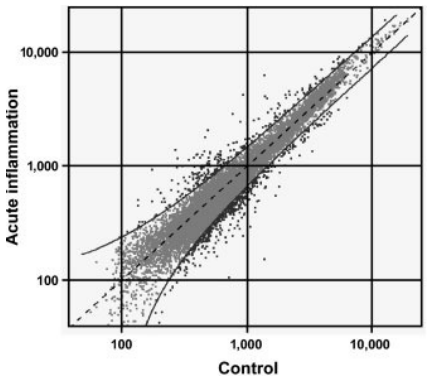


Fig. 2. Data normalization and CI for a significant variation of a given mRNA level. Total RNAs from two human liver samples were hybridized over an array and for every mRNA the two normalized signals were plotted (abcissa, control patient; ordinate, patient with an acute, systemic inflammation). The normalization results in most mRNA values were centered on the y = x axis (central, **dotted line**). The CI (**solid lines**, $\alpha = 0.05$) for nonsignificant fluctuations (grey squares) is inversely related to the absolute signal level and identifies the outliers (black squares) as inflammation-regulated mRNAs.

a given cDNA probe spotted over the entire array. This was verified over a large range of hybridization signals with various probes whose coefficient of variation usually was 10% to 12% (see supplementary Table s1A online (http://interscience.wiley.com/jpages/0270-9139/suppmat/index.html), as well as on our web site [www.lille.inserm.fr/u519/coulouarnetal03.html]). In addition, we verified (data not shown) that the genes and gene clusters highlighted in the current study were in no way associated with probes spatially clustered onto the filters or to local variations of the background that may increase the false discovery rate.[25] Third, many IMAGE clones are partial cDNAs or expressed sequence tags randomly located within a full-length cDNA, whereas our labeled targets primed with an anchored oligo(dT)VN primer preferentially covered the 3′ end of the cognate mRNA. Therefore, we verified that a difference in a given mRNA level between targets was consistantly detected whatever the location of a partial cDNA probe within the corresponding full-length cDNA (see our supplementary Table s1B online). Fourth, when comparing two complex targets in terms of induction/repression of a given mRNA level, relying on an arbitrary cutoff for a significant variation of the ratio (mRNA level in condition A / mRNA level in condition B) has been highly criticized.[26] To identify significantly different mRNA levels between samples, we used a statistically valid CI that varies with the absolute signal level (Fig. 2).

To store all the information and other data, such as clinical data, we developed the *LiverTools* database (Fig. 3). The data are entered into six sections that cover the MIAME recommendations.[23]

Tissue or Cell Type-Dependent Expression of Liver-Expressed Genes. *Liverpool* was probed with total RNA samples from various sources and the presence versus ab-

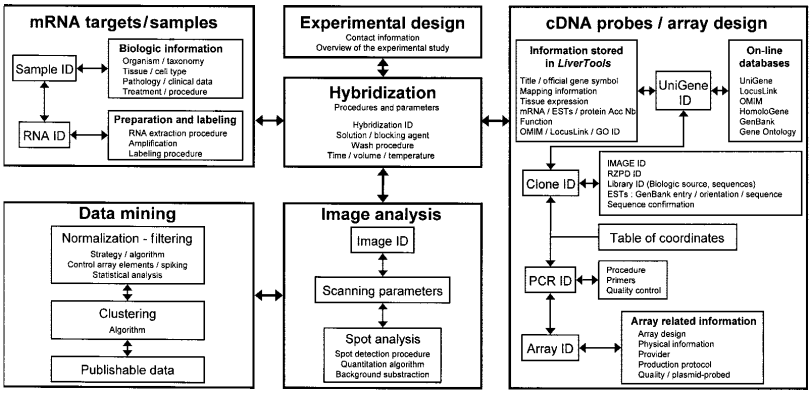


Fig. 3. *LiverTools*, a database tailored to liver transcriptome analysis. This database comprises six major sections, most of which are self-explanatory. The cDNA probes/array design section comprises the Hs. cluster and probe lists as well as other data. Some of the latter are entirely stored in *LiverTools* (*e.g.*, chromosomal location of the cognate gene, tissue-dependent gene expression, functional classification of the cognate protein according to the gene ontology consortium,[29] and related pathology as listed in the OMIM database [www.ncbi.nlm.nih.gov/omim/]). These data can be upgraded by direct links to on line databases. Other on line databases, such as HomoloGene, provide information to identify rodent orthologs. The various sections and available information comply with MIAME recommendations.[23] Unique identifiers (ID) are indicated whenever necessary. Most features are accessible upon request. Examples of queries that can be searched in *LiverTools* are posted on our web site (http://www.lille.inserm.fr/u519/coulouarnetal03.html).

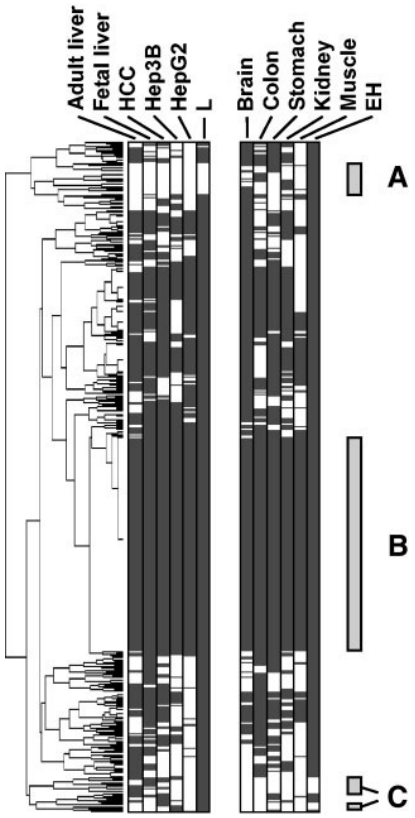


Fig. 4. Tissue-dependent expression of liver-expressed genes. The tissues listed from left to right are either liver related (**left panel**: HCC, hepatocellular carcinoma) or not liver related (**right panel**). The information in columns L or EH summarize the data for the various liver-related or extrahepatic tissue samples, respectively. Expression of a given mRNA in a given tissue sample is observed in an all-or-none fashion (**black horizontal line**) and a resulting hierarchical clustering is presented. Gray bars, mRNA subsets that are detected in all tissue samples analyzed but the liver (**A**), in all tissues (**B**), or in liver only (**C**). A random subset of only 1,000 mRNAs is shown for clarity.

sence of a positive signal was recorded for every gene tested. Overall, the probes for 8,921 Hs. clusters (90.5% of all tested clusters) provided a positive signal with at least one human tissue sample, whether liver or other organ related. As illustrated in Fig. 4, few probes (subset A, 1,359 Hs. clusters) did not provide a positive signal with at least one liver-related sample, a result that supports our liver-oriented probe selection. Most likely, the gene in subset A are expressed in the liver sample to an extent that is below our detection threshold. Another subset (subset C, 880 Hs. clusters) with a relatively limited size comprises genes that exhibit a liver-restricted expression and are mostly involved in amino acid and lipid metabolism, innate immunity, energy transformation, and detoxication. Finally, a major gene subset whose expression is found in every tissue sample tested (subset B) can be defined as being mostly composed of housekeeping genes. In agreement with this, subset B contains 38 known housekeeping genes.[18]

We also investigated whether the liver transcriptome exhibits any major difference when compared with the transcriptome of a single cell type, namely, the hepatocyte. When matching a crude human liver sample versus a hepatocyte-enriched fraction (<3% nonparenchymatous cells) purified from the same sample, we found that the homologous mRNAs from both sources were expressed to the same relative extent ($r = 0.92$, n = 12,493, $P < 10^{-4}$). This is consistent with the finding that the hepatocytes are mRNA-rich cells and account for 80% to 90% of the liver cell mass.[27]

***AP-Driven Genes in Human Liver* In Vivo.** We used *Liverpool* to gain a genomewide insight of the AP-associated events in the liver tissue samples obtained from patients experiencing AP. Based on the number of abnormalities within a set of eight biologic data measured at the time of surgery, the patients listed in Table 1 were given an AP score. They were divided into two subsets with a strong (patients 1–3) or moderate AP score (patients 4–6) and one subset of AP-free, control individuals (patients 7–10). Selecting genes whose mRNA level was abnormally high or low in at least two AP patients compared with its mean level in all four control patients, we identified 772 nonredundant genes. The overall data are presented as a bidimensional hierarchical clustering (Fig. 5A; the complete data set is available as Table s5 online). These data are reliable based on (1) the presence of numerous Hs. clusters corresponding to hepatic mRNAs whose levels are known to be either up-regulated by the AP such as C-reactive protein (CRP), orosomucoid, serum amyloid A (SAA) 1, fibrinogen α, β, and γ chains, phospholipase A2, cystathionase, annexin A1, the β chain of complement C1q (Clqβ), and the α chain of complement C4-binding protein, or down-regulated such as albumin (ALB), transferrin (TSF), transthyretin, alcohol dehydrogenase, and selenoprotein P[14] (also visit our web site for AP genes at www.lille.inserm.fr/u519/thematiques/equipe1/souryetal/index.html); (2) the unequal numbers of up-regulated and down-regulated genes (59% vs. 41%), in excellent agreement with earlier studies in humans and rats[28] (also visit our web site for AP genes); and (3) when applicable, the tight clustering of several cDNA probes corresponding to the same Hs. cluster as illustrated with complement C1qβ, SAA 1, TSF, or orosomucoid (Fig. 5A,B). As a quality control, Q-RT-PCR

Table 1. Biologic Data in Patients With an Acute, Systemic Inflammation and Control Patients

Patient No.	Sex	Age	Pathology*	Histology†	Fever	Leukocytes ($<10^4$/mm^3)	Hemoglobin (7.5-10 mmol/L)	CRP (<5 mg/L)‡	HAP (0.5-2.5 g/L)‡	ORM (0.5-1.2 g/L)‡	ALB (39-46 g/L)§	TSF (1.7-2.8 g/L)§	AP score‖
Acute inflammation													
1	M	53	Hepatic abcess + amibiasis	++	Yes	12,900	6.0	163	4.53	2.46	25.1	1.33	8 (+ high ESR)¶
2	M	76	Hepatic abcess + cholangiocarcinoma	+++	Yes	19,900	8.3	71.7	1.91	1.54	28.0	0.81	6 (+ high ESR)¶
3	F	66	Hepatic metastasis (ovary)	+	No	9,100	7.2	29.1	5.35	2.37	36.4	2.73	5 (+ high ESR)¶
4	F	28	Sclerosing cholangitis	++	No	12,300	9.1	24.6	2.13	1.20	62.8	1.65	3
5	M	52	Hepatic metastasis (bronchus)	0	No	7,800	9.4	19.7	2.09	1.32	45.0	2.01	2
6	F	65	Hepatic metastasis (colon)	++	No	10,400	9.1	12.1	1.92	1.19	43.8	2.90	2
Controls													
7	M	62	Hepatic metastasis (colon)	++	No	6,500	9.1	<5	1.44	0.75	40.2	2.28	0
8	M	76	Hepatic metastasis (colon)	+	No	5,530	8.8	<5	0.59	0.87	40.6	3.16	0
9	M	74	Hepatic metastasis (colon)	+	No	4,300	9.0	<5	0.67	0.94	41.4	1.90	0
10	F	94	Hepatic metastasis (breast)	++	No	Not done	Not done	<5	0.89	0.96	50.5	2.21	0

Abbreviations: CRP, C-reactive protein; HAP, haptoglobin; ORM, orosomucoid; ALB, albumin; TSF, transferrin; AP, acute phase; ESR, erythrocyte sedimentation rate; M, male; F, female.

* Whenever a liver was resected for metastasis, the tumor origin is noted in parentheses.

† Extent of local inflammation in liver, if any, as judged from the number and identity of white blood cells found in liver lobules and portal and sinusoid vessels: 0, none; + to +++, weak to strong.

‡ Plasma proteins up-regulated by the AP. Values are underlined when they are above the normal range.

§ Plasma proteins down-regulated by the AP. Values are underlined when they are below the normal range.

‖ Number of abnormal values as underlined in the columns "Fever" to "TSF."

¶ Patients 1 to 3 also had an abnormally high ESR but this information was not available for patients 4 to 6.

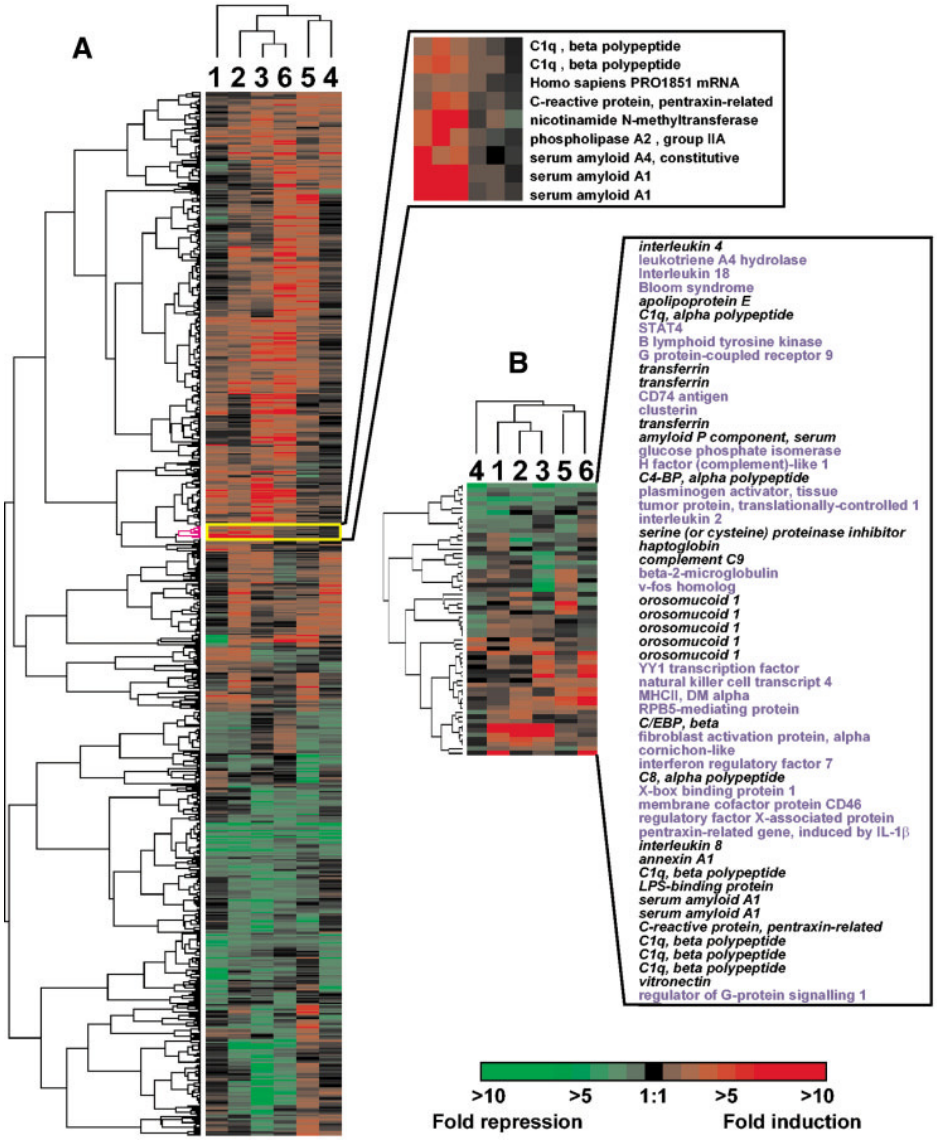


Fig. 5. Bidimensional hierarchical clustering of six AP patients and a set of AP-regulated mRNA levels in the liver. The AP patients (patients 1-6 in Table 1) are clustered horizontally. In every AP patient a change (fold) in a given mRNA level (average of three measures) is expressed with respect to the mean level in four control patients. This change is shown as a colored bar of variable intensity. The complete set of mRNA level changes is clustered vertically. (**A**) Shown are 772 mRNA levels corresponding to 699 Hs. clusters and 73 further cDNAs not listed as Hs. clusters (the complete list along with this figure is available on line). A given mRNA is shown whenever its hepatic level was significantly up-regulated or down-regulated in at least two AP patients. (**Window**) Tight clustering of mRNA changes identified by several probes corresponding to a single gene (*e.g.*, C1qβ and SAA 1). (**B**) Shown are 56 mRNAs and the corresponding 47 gene names. These genes are a subset of the 772 genes selected in **A** and belong to the anti-pathogen response gene category.[29] Genes that have long been known to be AP modulated are italicized whereas newly identified targets of the AP response are identified by blue block letters.

of several mRNAs whose levels measured by arrays exhibited extensive changes between our 10 liver samples was performed. In all instances, an excellent correlation ($P \leq .01$) of the values obtained by arrays versus Q-RT-PCR was found: CRP, $r = 0.90$; haptoglobin, $r = 0.76$; ALB, $r = 0.82$.

The unsupervised clustering based on the complete set of 772 genes identified three major patient groups (patient 1 vs. patients 2, 3, 6 vs. patients 4, 5) that did not match the moderate versus strong AP score (Fig. 5A). We then determined whether one or more functionally defined gene subsets would better cluster patients 1 to 3. Generally known protein functions allow the division of the *Liverpool* gene list into 42 main subsets[29] (see details in our supplementary Figure s1 online). Excluding, from the 772 genes identified above, those coding for orphan or putative proteins and next dividing the remaining 314 genes into functionally defined subsets resulted in a clear-cut clustering of patients 1 to 3 in one instance only, that is, when considering the so-called "anti-pathogen response"-associated genes (47 genes in our study; Fig. 5B). The anti-pathogen function immediately argues against a selection of this subset by chance. Some of the genes in this subset (italicized names in Fig. 5B) code for proinflammatory cytokines (*e.g.*, interleukin [IL]-8), TFs (*e.g.*, CCAAT-enhancer binding protein [C/EBP]-β), and several complement components and APPs (*e.g.*, CRP, SAA 1, orosomucoid) that are AP regulated.[14,30] Some other gene products have not been previously associated with AP (names written in blue block letters in Fig. 5B) and extend the number of AP-regulated anti-pathogen response proteins. These observations underscore that, in the human liver, the so-called anti-pathogen response proteins largely overlap with the AP-regulated proteins.

Novel Markers of AP Extent. Cumulated bibliographic data indicate that fewer than 200 AP-regulated genes have been identified in the liver (see our web site for AP genes). Therefore, many of the 772 genes shown in Fig. 5A have not been previously associated with the liver response to AP and provide novel potential markers for this condition. However, it may be argued that some of the mRNA regulations observed in our AP patients may not result from the inflammatory syndrome. Therefore, a positive or negative correlation between the extent of change in a given mRNA level and the extent of inflammation based on the AP score were used to rank the 772 genes. The resulting list of 154 genes for which such a correlation is statistically significant ($r > 0.63$ or $r < -0.63$, $P < .05$) is illustrated in Table 2 and is accessible as supplementary Table s2 online. It is noteworthy that the identity of all genes identified at this stage was further checked by resequencing the corresponding IMAGE clones used as probes.

Table 2. Correlation Between the Extent of AP and Various Hepatic mRNA Levels

mRNA	IMAGE Clone*	r†
Up-regulation		
Identified mRNA/acknowledged marker of AP (total : 16)		
C1q β	112128	0.93
SAA4	1917449	0.90
SAA1	161456	0.88
PLA2, group IIA	138064	0.80
CRP	121659	0.76
ORM	199253	0.70
Metallothionein 1G	194569	0.64
Identified mRNA/novel marker of AP (total : 30)		
Sequestosome 1	252234	0.86
Natural killer cell transcript 4	341021	0.79
TNF receptor-associated factor 5	145410	0.70
Cytoplasmic dynein, H1	122483	0.68
Insulin-like growth factor binding protein 2	78100	0.67
Sialyltransferase 9	781941	0.66
Putative mRNA/novel marker of AP (total : 26)		
ESTs	128768	0.90
Hypothetical protein MGC4840	280494	0.89
Down-regulation		
Identified mRNA/acknowledged marker of AP‡ (total : 4)		
c-Jun	321923	−0.98
Glucagon receptor	124201	−0.87
Di-carbonyl/L-xylulose reductase	758030	−0.69
IL-4	home PCR§	−0.68
Identified mRNA/novel marker of AP (total : 40)		
Down's syndrome CAM-like 1	2136882	−0.96
Protection of telomeres 1	52443	−0.94
MAP4K4	347368	−0.87
Transforming growth factor β3	796607	−0.85
Laminin, α5	770918	−0.83
Collagen IVα1	109703	−0.76
Putative mRNA/novel marker of AP (total : 38)		
ESTs	1841283	−0.92
KIAA1387 protein	504494	−0.91

NOTE. The complete set of data is available as supplementary Table s2 on line.

Abbreviations: AP, acute phase; mRNA, messenger RNA; CRP, C-reactive protein; ORM, orosomucoid; TNF, tumor necrosis factor; IL-4, interleukin-4.

*IMAGE clone number is used as the unique identifier of a given probe.

†With n = 10 (*i.e.*, patients 1-10), the correlation is statistically significant ($P < .05$) whenever $r > 0.63$ (up-regulated genes) or $r > -0.63$ (down-regulated genes).

‡The corresponding mRNA and/or protein previously have been shown to be regulated in acute inflammation or by proinflammatory cytokines (30, and our website for AP genes).

§An IMAGE clone was not available.

These 154 genes contain a subset of 20 genes that have long been known to be AP regulated. Of these 20 mRNAs, 12 code for positive APPs (60%) and display an excellent correlation with the AP score (*e.g.*, C1qβ, SAA 4, CRP; $r \geq 0.70$). In contrast, only 4 of these 20 mRNAs are down-regulated and only one codes for a plasma protein (IL-4). Strikingly, none of the mRNAs for well-known negative APPs (*e.g.*, ALB, TSF, transthyretin) was found.

These 154 genes further contain a set of 134 novel genes. As illustrated in Table 2, this set includes 70 known genes whose expression had not yet been shown to be AP modulated, as well as 64 genes for putative proteins. Many of these 134 novel mRNAs, most of them down regulated, have a correlation value greater than ± 0.8. Therefore, they are excellent candidates as novel markers of clinical interest. We then determined whether some of the corresponding proteins are secreted as they would lend themselves to quantitation in body fluids for diagnosis and prognosis purposes. To address this, we took advantage of another study in which protein secretion was suggested from a combination of ontology and sequence-based analyses.[31] Matching these data with our 134 AP-regulated mRNAs identified five mRNAs coding for secreted proteins (*e.g.*, natural killer cell transcript 4 [NK4], insulin-like growth factor binding protein 2 [IGFBP2], transforming growth factor-β3, lamininα5, and collagenIVα1).

Hepatocyte Origin and Cytokine-Regulated Abundance of the Novel mRNAs. We verified that the 134 newly identified mRNAs were synthesized in hepatocytes in a cytokine-dependent manner. This was proven by measuring some of these mRNAs in human Hep3B hepatoma cells stimulated *in vitro* with a proinflammatory, cytokine-enriched conditioned medium versus a negative control medium. This approach was proven to be reliable by the expected up-regulation of the CRP mRNA level and the down-regulation of the ALB mRNA level after 6 or 16 hours of stimulation (Fig. 6, first 2 panels). Moreover, the levels of the six other mRNAs of interest (NK4, IGFBP2, nuclear receptor subfamily 1, group 2, member 2 [NR1D2], protection of telomeres 1, [POT1], MAP4K4, CGI-41) were up-regulated or down-regulated in the conditioned medium-challenged Hep3B cells (Fig. 6), in agreement with the data obtained for the patients in the current study. Only one mRNA (tumor necrosis factor-associated factor 5 [TRAF5]) did not exhibit a significant modulation in Hep3B cells, which may result from the limited time frame of the stimulation used.

Inflammation-Regulated Sites in the Promoters of the Novel AP-Regulated Genes. We investigated whether the promoters of the 154 genes whose mRNA levels strongly correlate with the extent of AP exhibit a high frequency of binding sites for AP-regulated TFs. In the liver, these TFs mostly comprise nuclear factor κB (NFκB), activating protein-1, signal transducer and activator of transcription-3 (STAT-3), and some members of the C/EBP family.[14] The occurrence of potential binding sites for one or more of these factors in the promoters of some of the 154 genes was compared with their occurence in a random set of control genes transcribed at least in the liver. The results detailed in our supplementary Table s3 point to a significantly higher number of potential sites for NFκB and activating protein-1 in the AP-regulated genes compared with the control gene set. We conclude that the majority of the 154 known or orphan genes (Tables 2 and s2) are likely to be controlled by one or more AP-regulated TFs.

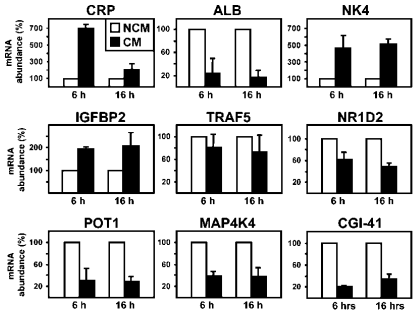


Fig. 6. Proinflammatory cytokine-dependent regulation of mRNA levels in Hep3B hepatoma cells. Hep3B cells were stimulated for 6 or 16 hours with a cytokine-enriched, conditioned medium (CM) or a negative, control medium (NCM). Total RNA samples were used for determination of specific mRNA levels by Q-RT-PCR (the oligo sequences are detailed on line). The values were normalized with the level of glyceraldehyde-3-phosphate dehydrogenase mRNA that is not modulated by proinflammatory cytokines.[50] CRP and ALB were used as controls for an up-regulated or down-regulated mRNA sample, respectively. Other mRNA samples code for NK4, IGFBP2, TRAF5, NR1D2, protection of telomeres 1 (POT1), MAP4K4, and the putative protein CGI-41. Data are mean $\pm$ SD from three independent cultures and is expressed as a percentage of the mRNA level in cells stimulated with the NCM (100%).

The Tightly Controlled mRNA Levels Cover From Signaling to Positive APPs. The 154 mRNAs listed in Tables 2 and s2 code for proteins covering prominent functions that are detailed in Table s2. This includes membrane/cytoskeleton organization, protein sorting and secretion, general metabolism, and detoxication. Two down-regulated genes participate in proteolysis and two up-regulated, early response genes protect against proteasome-mediated proteolysis. Other mRNAs code for secreted proteins (*e.g.*, extracellular matrix components; cytokines). Most importantly, four other mRNA subsets are of immediate relevance in an AP context.

One subset codes for TRAF5, calpain 6, MAP4K4, and chemokine-like factor superfamily 6 that all play critical functions at early stages of AP-driven signaling. Specifically, TRAF5 (up-regulated) is an accessory molecule for the tumor necrosis factor-α receptor and is involved in activation of the NFκB pathway. In addition, both calpain 6 and MAP4K4 control the tumor necrosis factor-α/IL-1–activated MAP kinase pathway. Calpains control a protein tyrosine kinase 2-mediated cascade that targets

the MAP2K1 protein. MAP4K4 is a serine/threonine kinase that specifically activates c-Jun kinase 1. Its down-regulation, as observed in the current study, limits c-Jun activity and, therefore, is in keeping with the lowered c-Jun mRNA abundance that we observed.

Another subset is involved in transcription and indicates a trend towards transcriptional limitation. Down-regulated mRNAs for activators include c-Jun, c-Myc, and the structure-specific recognition protein 1 that is a p63 coactivator. Other mRNAs correspond to factors involved in transcriptional repression (prospero-related homeobox 1, RPB5-mediating protein, rev-Erb-related NR1D2, Ets variant Etv3) or chromatin rearrangement (MSL3-like1, DNA helicase type 2, structure-specific recognition protein 1). It is noteworthy that none of the mRNAs for transcriptional activators that are known to be up-regulated by the AP in liver, namely, NFκB, C/EBP-β and C/EBP-δ, and STAT-3,[14] was found in this list. These mRNAs are up-regulated only after latent NFκB, C/EBP-β, and STAT-3 molecules that preexist in the cytosol of quiescent hepatocytes have been rapidly imported to the nucleus at the onset of AP. Remarkably, our list of up-regulated mRNAs included RAN-binding protein 16, which participates in such a nuclear protein import.

A third subset participates in N-glycan maturation and processing, including the controls of sialic acid addition by sialic acid synthase and ganglioside synthesis by sialyltransferase 9. Sialylation/desialylation of glycoproteins is critical in self-recognition and intercellular communication in innate immunity[32] and the sialic acid residues present on gangliosides participate in activation of the AP-triggered JAK-STAT pathway.[33]

Finally, the mRNAs that are most tightly correlated with the AP score preferentially code for up-regulated plasma APPs. The latter amount to 15 of 30 (50%) of the functionally identified mRNAs whose correlation is greater than or equal to +0.70.

Discussion

The liver transcriptome in primates or rodents has been increasingly studied using array technology.[11,12] A limited probe diversity (<2,000 genes) prevented complete transcriptome coverage in approximately 50% of the studies. Current estimates of the number of genes expressed in any given vertebrate tissue sample, including the liver, range from 10,000 to 15,000.[12,34] Other studies of the liver transcriptome were conducted using probes for 4,000 to 23,000 genes that were not selected with respect to this organ. This resulted in only 2,500 to 4,500 informative genes[35–37] or in a figure that is not publicly available.[28,38–45] As yet, only one study has used 11,000 probes that were specifically tailored to liver transcriptome analysis. However, this analysis resulted in only 30% to 67% informative probes, depending on the target mRNA mixture.[46] Therefore, the current study is the first to cover most of the liver transcriptome in humans as judged from the high rate of detected mRNAs (86% of 9,858 Hs. clusters) as well as the relatively high number of mRNAs transcribed by subset C of liver-specific genes. Subset C comprises a limited size compared with the total number of genes transcribed in the liver, which supports earlier observations.[34,41,47,48] However, subset C contains 880 Hs. clusters, which is far more than the number of liver-specific genes identified in a study performed with only 10,000 unselected genes.[34] Along with comprehensive coverage of the liver transcriptome owing to *Liverpool*, we have conjointly developed *LiverTools*. We believe that such a combination dedicated to liver transcriptome analysis has no counterpart.

We focused on the liver response to acute, systemic inflammation because this condition strongly regulates numerous genes in the liver. Therefore, deciphering the underlying mechanisms is of interest for our understanding of liver biology as well as for clinical purposes. This is the first study of the AP-dependent modulations of the liver transcriptome in humans *in vivo* that is amenable to statistical analysis. It could be argued that in some patients the altered abundance of some mRNAs resulted from the underlying cancerous disease and was inappropriately ascribed to the AP. However, this is unlikely given the heterogeneity of the cancers involved, the presence of liver samples from noncancerous patients, the correlation of mRNA abundance with the inflammation index, and our controls with cytokine-challenged cell cultures. The data obtained for our patients and in cytokine-challenged Hep3B cells and our search for enrichment in binding sites for AP-regulated TFs in the liver indicate that the current study deals mostly, if not exclusively, with the hepatocyte transcriptome. This is in keeping with our observation that the mRNA levels in whole liver and isolated hepatocytes are strongly correlated. However, our probe selection was made, at least partly, from whole liver cDNA libraries and, therefore, lends itself to transcriptome analysis in nonparenchymatous liver cells as well.

A major finding of the current study is the selection of 20 known and 134 novel human mRNAs whose hepatic level significantly correlates with the extent of AP. This correlation value allowed us to rank the resulting series of 134 newly identified mRNAs as novel AP markers and, in this respect, some of them outperformed some well-known markers. Remarkably, haptoglobin, ALB, and TSF are classically used as AP markers[14,15] and they participated in the initial assignment of an AP score to our

patients. Haptoglobin, ALB, and TSF mRNAs were also listed in our preliminary selection of 772 genes but they did not pass our final selection for genes whose mRNA abundance correlates with the AP score. A high level of albuminemia or transferrinemia observed in patients 3, 4, and 6 in Table 1 supports this finding. We conclude that markers such as haptoglobin, ALB, and TSF are of interest in the detection of an AP but they poorly perform beyond their use in an all-or-none fashion (as in Table 1). On the contrary, some novel genes also code for secreted proteins (collagenIVα1, IGFBP2, lamininα5, NK4, transforming growth factor-β3) but with the added bonus of an up-regulated or down-regulated mRNA level that strongly correlates with the AP score. The levels of the corresponding proteins deserve extensive studies in body fluid samples. They will help solve the current need for novel APPs as sensitive diagnosis and prognosis tools.[15] It will also be worth investigating to which extent the mRNA versus protein levels correlate. In this respect, the plasma protein versus hepatic mRNA correlation was searched for all APPs listed in Table 1 and was found to be highly variable between APPs ($0.2 < r < 0.8$), in agreement with other studies.[49]

One could have expected that the abundance of the mRNAs that are controlled most directly by cytokine-triggered receptors and associated cascades (*e.g.*, some mRNAs for TFs) would best correlate with the AP extent. However, the functions of the proteins coded by the 154 mRNA samples in Table 2 indicate that the liver is able to adapt its response to the AP extent by virtue of a fine-tuned change in abundance of functionally diverse mRNAs. This is well illustrated by four prominent mRNA sets that code for early signaling molecules TFs, N-glycosylation enzymes, or positive APPs. These four sets summarize the exquisite precision of the AP-driven hepatocyte response. The liver is able to follow the extent of AP owing to a fine tuning of some mRNA levels controlling most, if not all, intracellular events from early signaling to the final secretion of proteins involved in innate immunity.

Acknowledgment: Dr. A. Vente (Deutsches Ressourcenzentrum für Genomforschung [RZPD], Berlin) is acknowledged for his assistance in the early stage of IMAGE clone selection. The authors are indebted to Dr. G.M. Hampton for providing a list of secreted proteins, to Dr. S. Claeyssens for plasma protein determinations, and to M. Hiron for excellent technical assistance. The assistance of C. Bansard, F. Caillot, and G. Saint-Auret in clone resequencing is appreciated. The authors also thank Prof. R.P. Erickson for a critical reading of the manuscript.

References

1. Aravind L. Guilt by association: contextual information in genome analysis. Genome Res 2000;10:1074–1077.
2. Vidal M. A biological atlas of functional maps. Cell 2001;104:333–339.
3. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J Mol Biol 2001; 314:1053–1066.
4. Holloway AJ, van Laar RK, Tothill RW, Bowtell DDL. Options available—from start to finish—for obtaining data from DNA microarrays II. Nat Genet 2002;32(Suppl):481–489.
5. Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, Kay SA, et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 2001;106:413–415.
6. Wong GKS, Passey DA, Yu J. Most of the human genome is transcribed. Genome Res 2001;11:1975–1977.
7. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001;2:418–427.
8. Wright AF, van Heyningen V. Short cut to disease genes. Nature 2001; 414:705–706.
9. Arias IM, Boyer JL, Fausto N, Jakoby WB, Schachter DA, Shafritz DA. The liver. Biology and Pathobiology. 3rd ed. New York: Raven Press, 1994.
10. MacSween RNM, Burt AD, Portmann BC, Ishak KG, Scheuer PJ, Anthony PP. Pathology of the Liver. 4th ed. Orlando: Harcourt Brace, 2001.
11. Craig Barton M, Stivers DN. Microarray analysis of hepatic-regulated gene expression: specific applications and nonspecific problems. HEPATOLOGY 2002;35:727–729.
12. Shackel NA, Gorrell MD, McCaughan GW. Gene array analysis and the liver. HEPATOLOGY 2002;36:1313–1325.
13. Olivier E, Soury E, Risler JL, Smih F, Schneider K, Lochner K, Jouzeau JY, et al. A novel set of hepatic mRNAs preferentially expressed during an acute inflammation in rat represents mostly intracellular proteins. Genomics 1999;57:352–364.
14. Ruminy P, Gangneux C, Claeyssens S, Scotte M, Daveau M, Salier JP. Gene transcription in hepatocyte during the acute phase of a systemic inflammation: from transcription factors to target genes. Inflamm Res 2001;50:383–390.
15. Anderson NL, Anderson NG. The human plasma proteome. History, character, and diagnostic prospects. Mol Cell Proteomics 2002;1:845–867.
16. Yoo JY, Desiderio S. Innate and acquired immunity intersect in a global view of the acute-phase response. Proc Natl Acad Sci U S A 2003;100: 1157–1162.
17. Masson S, Daveau M, Francois A, Bodenant C, Hiron M, Teniere P, Salier JP, et al. Up-regulated expression of HGF in rat liver cells after experimental endotoxemia: a potential pathway for enhancement of liver regeneration. Growth Factors 2001;18:237–250.
18. Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. Comparison of human adult and fetal expression and identification of 535 housekeeping/ maintenance genes. Physiol Genomics 2000;2:143–147.
19. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, et al. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 2000;28:316–319.
20. Novak JP, Sladek R, Hudson TJ. Characterization of variability in large-scale gene expression data: implications for study design. Genomics 2002; 79:104–113.
21. Ihaka R, Gentleman RR. A language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.
22. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–14868.
23. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 2001;29: 365–371.
24. Lercher MJ, Urrutia AO, Hurst LD. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 2002;31:180–183.
25. Colantuoni C, Henry G, Zeger S, Pevsner J. Local mean normalization of microarray element signal intensities across an array surface: quality control

and correction of spatially systematic artifacts. BioTechniques 2002;32: 1316–1320.
26. Firestein GS, Pisetsky DS. DNA microarrays: boundless technology or bound by technology? Guidelines for studies using microarray technology. Arthritis Rheum 2002;46: 859–861.
27. Harbrecht BG, Billiar TR, Curran RD. Experimental models for studying the interaction of Kuppfer cells and hepatocytes. In: Billiar TR, Curran RD, eds. Hepatocyte and Kupffer Cell Interactions. Boca Raton, FL: CRC Press, 1992:55–70.
28. Chinnaiyan AM, Huber-Lang M, Kumar-Sinha C, Barrette TR, Shankar-Sinha S, Sarma VJ, Padgaonkar VA, et al. Molecular signatures of sepsis: multiorgan gene expression profiles of systemic inflammation. Am J Pathol 2001;159:1199–1209.
29. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–29.
30. Soury E, Olivier E, Simon D, Ruminy P, Kitada K, Hiron M, Daveau M, et al. Chromosomal assignments of mammalian genes with an acute inflammation-regulated expression in liver. Immunogenetics 2001;53:634–642.
31. Welsh JB, Sapinoso LM, Kern SG, Brown DA, Liu T, Bauskin AR, Ward RL, et al Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. Proc Natl Acad Sci U S A 2003;100: 3410–3415.
32. Kim OS, Park EJ, Joe EH, Jou I. JAK-STAT signalling mediates gangliosides-induced inflammatory response in brain microglial cells. J Biol Chem 2002;277:40594–40601.
33. Medzhitov R, Janeway CA Jr. Decoding the patterns of self and nonself by the innate immune system. Science 2002;296:298–300.
34. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, et al. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 2002;99:4465–4470.
35. Bigger CB, Brasky KM, Lanford RE. DNA microarray analysis of chimpanzee liver during acute resolving hepatitis C virus infection. J Virol 2001;75:7059–7066.
36. Okabe H, Satoh S, Kato T, Kitahara O, Yanagawa R, Yamaoka Y, Tsunoda T, et al. Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res 2001;61:2129–2137.
37. Yano N, Habib NA, Fadden KJ, Yamashita H, Mitry R, Jauregui H, Kane A, et al. Profiling the adult human liver transcriptome: analysis by cDNA array hybridization. J Hepatol 2001;35:178–186.
38. Cao SX, Dhahbi JM, Mote PL, Spindler SR. Genomic profiling of short- and long-term caloric restriction effets in the liver of aging mice. Proc Natl Acad Sci U S A 2001;98:10630–10635.
39. Graveel CR, Jatkoe T, Madore SJ, Holt AL, Farnham PJ. Expression profiling and identification of novel genes in hepatocellular carcinomas. Oncogene 2001;20:2704–2712.
40. Shih DQ, Bussen M, Sehayek E, Ananthanarayanan M, Shneider BL, Suchy FJ, Shefer S, et al. Hepatocyte nuclear factor-1 alpha is an essential regulator of bile acid and plasma cholesterol metabolism. Nat Genet 2001; 27:375–382.
41. Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, et al. Gene expression patterns in human liver cancers. Mol Biol Cell 2002;13:1929–1939.
42. Iizuka N, Oka M, Yamada-Okabe H, Mori N, Tamesa T, Okada T, Takemoto N, et al. Comparison of gene expression profiles between hepatitis B virus- and hepatitis C virus-infected hepatocellular carcinoma by oligonucleotide microarray data on the basis of a supervised learning method. Cancer Res 2002;62:3939–3944.
43. Kelley-Loughnane N, Sabla GE, Ley-Ebert C, Aronow BJ, Bezerra JA. Independent and overlapping transcriptional activation during liver development and regeneration in mice. HEPATOLOGY 2002;35:525–534.
44. Chuma M, Sakamoto M, Yamazaki K, Ohta T, Ohki M, Asaka M, Hirohashi S. Expression profiling in multistage hepatocarcinogenesis: identification of HSP70 as a molecular marker of early hepatocellular carcinoma. HEPATOLOGY 2003;37:198–207.
45. Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, Simon R, et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 2003;9:416–423.
46. Xu L, Hui L, Wang S, Gong J, Jin Y, Wang Y, Ji Y, et al. Expression profiling suggested a regulatory role of liver-enriched transcription factors in human hepatocellular carcinoma. Cancer Res 2001;61:3176–3181.
47. Eickhoff H, Schuchhardt J, Ivanov I, Meier-Ewert S, O'Brien J, Malik A, Tandon N, et al. Tissue gene expression analysis using arrayed normalized cDNA libraries. Genome Res 2000;10:1230–1240.
48. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, et al. A compendium of gene expression in normal human tissues. Physiol Genomics 2001;7:97–104.
49. Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL, Giordano TJ, et al. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 2002;1:304–313.
50. Essani NA, McGuire GM, Manning AM, Jaeschke H. Endotoxin-induced activation of the nuclear transcription factor kappa B and expression of E-selectin messenger RNA in hepatocytes, Kupffer cells, and endothelial cells in vivo. J Immunol 1996;156: 2956–2963.

Chapitre 5

Une cinétique de la phase aiguë in vitro

La phase aiguë de la réponse inflammatoire est coordonnée par un grand nombre de médiateurs dont notament les cytokines pro-inflammatoires tels le *Tumor Necrosis Factor*(TNF)-α et l'interleukine (IL)-1β, principalement produits par les macrophages.
Ces deux cytokines induisent alors une seconde vague de cytokines telle l'IL-6 principalement par l'intermédiaire des fibroblastes et des macrophages. L'IL-6 amplifie la réponse de certains organes mais réprime également la production du TNFα et facilite de ce fait le retour à l'homéostasie.

La régulation des gènes dans le foie est fondamentale lors de la phase aiguë de l'inflammation. La phase aiguë régule notament nombre de gènes exprimés dans le foie et impliqués dans l'immunité innée et codant pour des protéines intracellulaires (Acute Phase-Regulated Intracellular Proteins, APRIP) et plasmatiques (Acute Phase Proteins, APP).
Ces régulations impliquent des altérations au niveau transcriptionnel et post-transcriptionnel et notament en terme d'abondance des ARNm, de leur stabilité ainsi que de leur traduction.
Pour autant l'importance de chacun de ces leviers de régulations reste à déterminer.

Bien que notre précédent travail ait permis de mettre en lumière de nouvelles APP et APRIP dont les modifications d'abondance sont étroitement corrélées *in vivo* au niveau d'inflammation sytémique aiguë, il ne permet cependant pas de déterminer les vagues cinétiques de régulations qui permettrait de mieux comprendre les interactions entre ces différents leviers.

Nous avons pour cela proposé un protocole expérimental permettant une analyse cinétique des altérations du transcriptome hépatique *in vitro*. À cette fin, des hépatocytes humains ont été stimulées à différents temps par des cytokines pro-inflammatoires, nous ouvrant ainsi une fenêtre sur la cinétique de la phase aiguë de l'inflammation.

Genome-Wide Response of the Human Hep3B Hepatoma Cell to Proinflammatory Cytokines, From Transcription to Translation

Cédric Coulouarn, Grégory Lefebvre, Romain Daveau, Franck Letellier, Martine Hiron, Laurent Drouot, Maryvonne Daveau, and Jean-Philippe Salier

Given the unknown timing of the onset of an acute systemic inflammation in humans, the fine tuning of cascades and pathways involved in the associated hepatocyte response cannot be appraised *in vivo*. Therefore, the authors used a genome-wide and kinetic analysis in the human Hep3B hepatoma cell line challenged with a conditioned medium from bacterial lipopolysaccharide-stimulated macrophages. A complete coverage of the liver transcriptome disclosed 648 mRNAs whose change in abundance allowed for their clustering in mRNA subsets with an early, intermediate, or late regulation. The contribution of transcription, stability, or translation was appraised with genome-wide studies of the changes in nuclear primary transcripts, mRNA decay, or polysome-associated mRNAs. A predominance of mRNAs with decreased stability and the fact that translation alone controls a significant number of acute phase–associated proteins are prominent findings. Transcription and stability act independently or, more rarely, cooperate or even counteract in a gene-by-gene manner, which results in a unidirectional change in mRNA abundance. Waves of mRNAs for groups of functionally related proteins are up- or downregulated in an ordered fashion. This includes an early regulation of transcription-associated proteins, an intermediate repression of detoxication and metabolism proteins, and finally an enhanced translation and transport of a number of membranous or secreted proteins along with an enhanced protein degradation. In conclusion, this study provides a comprehensive and simultaneous overview of events in the human hepatocyte during the inflammatory acute phase. *Supplementary material for this article can be found on the HEPATOLOGY website (http://www.interscience.wiley.com/jpages/0270-9139/suppmat/index.html).* (HEPATOLOGY 2005;42:946-955.)

The acute phase (AP) of the inflammatory response is coordinated by a large number of mediators, such as the pro-inflammatory cytokines tumor necrosis factor (TNF)-α and interleukin (IL)-1β, mainly produced by macrophages.[1] TNF-α and IL-1β then promote a second wave of cytokines, such as IL-6, mostly released by macrophages and fibroblasts.[1,2] IL-6 is a dual, pro- and anti-inflammatory cytokine. It amplifies the response of various organs to AP while it downregulates the production of TNF-α, thereby facilitating the so-called resolution phase and a return to homeostasis.[1-4]

Altered gene regulation in the liver is a hallmark of AP.[1,2] Specifically, the AP regulates many liver-expressed genes involved in innate immunity and coding for AP-regulated intracellular proteins (APRIPs) and plasma acute phase proteins (APPs), which are transiently up- or downregulated and consequently classified as positive or negative APRIPs/APPs.[5-8] These regulations entail transcriptional or post-transcriptional step(s), which results in altered mRNA abundance, stability, or translation,[6,9] but the relative importance of these controls remains to be assessed. Transcriptome studies in rodents have partly dissected the elaborate and dynamic process that takes place in the liver in the course of AP.[10,11] In contrast, a genome-

Abbreviations: AP, acute phase; TNF, tumor necrosis factor; IL, interleukin; APRIP, acute phase-regulated intracellular protein; APP, acute phase protein; CM, conditioned medium; NCM, nonconditioned medium; q-RT-PCR, quantitative reverse transcription polymerase chain reaction; ARE, AU-rich element; MAO-B, monoamine oxidase B; NF-κB, nuclear factor κB; POLR-2F, polymerase (RNA) polypeptide 2F; CAM-1, calmodulin 1.

From INSERM U519 and IFRMP, Rouen, France.

Received March 1, 2005; accepted July 13, 2005.

C.C. and G.L. are the recipients of a fellowship from the French Ministry for Research and C.C. is the recipient of a fellowship from Association de Recherche sur le Cancer. Supported in part by grants from Association de Recherche sur le Cancer and Ligue contre le Cancer (J.-P.S.).

Address reprint requests to: J.P. Salier, INSERM U519, Faculté de Médecine-Pharmacie, 22 Bvd Gambetta, 76183 Rouen cedex, France. E-mail: Jean-Philippe.Salier@univ-rouen.fr; fax: (33) 235-14-85-41.

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI 10.1002/hep.20848

Potential conflict of interest: Nothing to report.

wide and kinetic view of the AP-induced changes in the human liver is not available. Owing to a coverage of the whole human liver transcriptome with a dedicated microarray, we recently identified the APRIP and APP mRNAs whose altered abundance best correlates with the extent of an acute, systemic inflammation *in vivo*.[12] However, deciphering the complete regulatory steps of signaling cascades and pathways involved in the response of the human liver to inflammation cannot be gained from studies *in vivo* that lack essential information such as the time of AP onset. Therefore, an analysis of the kinetics of transcriptome alteration in the human hepatocyte challenged the pro-inflammatory cytokines *in vitro* should provide a privileged window on the liver response to AP. With this approach, we have now observed that several waves of mRNAs for groups of functionally related proteins are up- or downregulated in an ordered fashion. Analysis of mRNA transcription, stability, and translation further indicated that in most instances these control steps act independently or, more rarely, cooperate or even counteract in a gene-by-gene manner, which still results in a unidirectional change in mRNA abundance.

Materials and Methods

Stimulation of a Hepatoma Cell Line With Pro-inflammatory Cytokines. The human Hep3B hepatoma cells (ATCC HB-8064) plated at 33% confluence were cultured for 48 hours, and the culture medium was next changed for a mixture made of a serum-free medium added with 20% (vol/vol) stimulated-macrophage conditioned medium (CM) enriched in TNF-α, IL-1β, IL-6, and IL-8 or nonconditioned medium (NCM) used as a control.[12] Paired cultures were challenged with CM versus NCM for a given length of time in 3 (time-course) or 2 (stability, transcription, or translation) independent experiments.

Determination of mRNA Abundance by Microarray or Polymerase Chain Reaction. Total RNAs were labeled and hybridized to our "Liverpool" microarray, which provides complete coverage of the human liver transcriptome (approximately 10,000 genes).[12] Quantitative reverse transcription polymerase chain reaction (q-RT-PCR) of mRNAs with the primers listed in our Supplementary Table 1 was done as described.[12]

Analysis of mRNA Stability by Actinomycin D and Microarray. The cells were stimulated with CM or NCM for a fixed time. The medium was replaced by serum-free medium containing 10 μg/mL actinomycin D (Sigma, St. Louis, MO), the dishes were kept at 37°C for 0, 15, 60, or 240 minutes (1 dish per time) and the resulting RNAs were labeled and hybridized to our microarray. Every mRNA 3′ untranslated region (3′UTR) was retrieved from the ENSEMBL data library, and a search for AU-rich elements (AREs) was made with a series of 30 published AREs,[13,14] owing to a locally developed PERL script.

Analysis of RNA Transcription by Run-on and Microarray. Nuclear primary transcripts labeled with [α-^{32}P]UTP by run-on assay were prepared as described by in Daveau et al.[15] and hybridized to our microarray for 64 hours followed by autoradiography for 1 week.

Analysis of mRNA Translation by Polysome Isolation and Microarray. Polysome fractionation was done essentially as described elsewhere.[16-18] The polysome-free or polysome-enriched fractions were pooled separately and then labeled and hybridized to our microarray.

Microarray Data Handling and Mining. Our general procedures for data handling were detailed previously.[12] For every detected mRNA, the normalized paired values obtained under NCM versus CM at a given time were considered to be significantly induced or repressed (folds) when their difference was outside a funnel-shaped confidence interval ($P < .05$) calculated from every mRNA detected within the experiment.[12] In time-course experiments, an mRNA abundance was considered to be CM-regulated at a given time point whenever a significant induction or repression occurred in at least 2 of 3 independent experiments. For mRNA stability, run-on, or polysome-related experiments, the mean values measured at a given time were used for determination of confidence intervals, from which outlier transcripts were considered to be significantly regulated (further specific calculations are provided in the legends of the tables and figures). K-means clustering was done with the Genesis software.[19] Protein functions and groups of functionally related mRNAs were based on the Gene Ontology Consortium.[20]

Protein Electrophoresis and Immunodetection. SDS-PAGE and immunodetection were performed as described.[21] Goat antibodies against monoamine oxidase B (MAO-B) (catalogue ref. sc-18401) or calmodulin 1 (CAM-1) (sc-1989) and mouse antibodies against RNA polymerase II (DNA-directed) polypeptide F (POLR-2F) (sc-21752) were from Santa Cruz Biotechnology (Santa Cruz, CA). Alexa Fluor 680–labeled rabbit anti-goat or anti-mouse IgGs used as a secondary antibody were from Molecular Probes (Eugene, OR). Fluorescent protein bands were quantified with the Odyssey Imaging System from Li-Cor (Lincoln, NE).

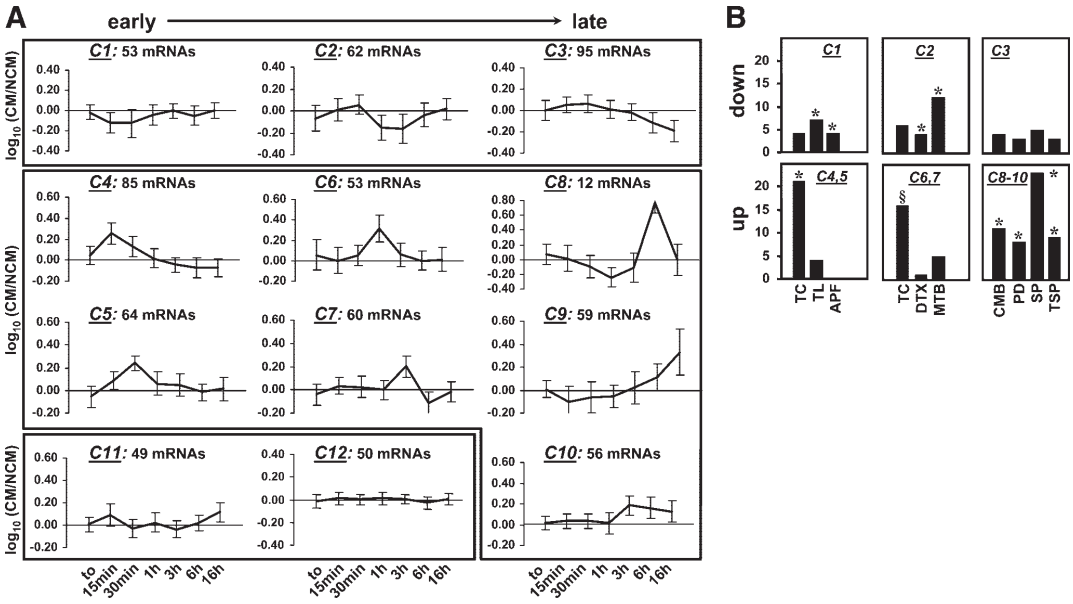


Fig. 1. Clustering and functions of Hep3B mRNAs whose abundance is regulated by a pro-inflammatory cytokine challenge. A total of 648 mRNAs whose abundance was found to be changed in CM- versus NCM-challenged cells were separated in 11 clusters C1 to C11 by k-means clustering. (A) Within each cluster, the total number of mRNAs is noted on top, the time points are noted on the abcissa, and the log ratio of mRNA abundance found in CM- versus NCM-challenged cells is depicted on the ordinate. The plot (solid line) is the mean ratio of abundance for all mRNAs included in this cluster and the vertical bar at any time point is the standard error of the mean. Upper solid frame: 3 clusters of downregulated mRNAs; medium solid frame: 7 clusters of upregulated mRNAs (note that C10 exhibits a change of the mean ratio at 3 consecutive time points); lower solid frame: a cluster of CM-regulated mRNAs lacking a prominent change at a given time (C11), and a cluster of non-regulated mRNAs used as negative controls (C12). (B) Prominent functions of proteins encoded by mRNA subsets. The clusters C1 to C10 above were gathered in broader clusters whenever necessary. Significance of under- or over-representation of mRNAs for a functional group by chi-square test: §$P < .05$; *$P < .01$. CM, conditioned medium; NCM, nonconditioned medium; TC, transcription; TL, translation; APF, apoptosis and proliferation; DTX, detoxication; MTB, metabolism; CMB, cell membrane; PD, protein degradation; SP, secreted proteins; TSP, transport.

Results

Kinetics of Cytokine-Induced Changes in mRNA Abundance. A time-course (0, 15, 30 minutes; 1, 3, 6, or 16 hours) of mRNA abundance changes was studied in the human Hep3B hepatoma cell line challenged with a pro-inflammatory cytokine-enriched CM versus control NCM. Our "Liverpool" microarray was used to identify every mRNA whose abundance exhibited a statistically significant difference under CM challenge at one or more given time points, which resulted in a selection of 648 such mRNAs, referred to as the Hep3B/CM mRNAs. To identify subsets of mRNAs with a similar timewise regulation of abundance, these Hep3B/CM mRNAs were next separated into clusters by k-means clustering. The latter is an unsupervised procedure that requires the number of clusters to be chosen beforehand.[19] We found that 11 clusters C1 to C11 conveyed appropriate information; all but 1 (C11) presented a typical up or down and time-dependent pattern (Fig. 1A). The complete list of mRNAs within each cluster is provided as Supplementary Table 2. C1 to C10 correspond to a down- (C1-C3) or upregulated abundance (C4-C10). Moreover, an early (<1 hour) change in C1, C4, and C5, an intermediate (1-3 hours) change in C2, C6, and C7, and a late (>6 hours) change in C3 and C8-C10 point to early or late CM-responsive genes. The mRNAs with an increased abundance predominated within the entire subset of early genes C1, 4, 5 (149 of 202 mRNAs, 73.8%). This feature was also found, albeit to a lower extent, in the two subsets of intermediate or late genes C2, 6, 7 (64.6%) and C3, 8, 9, 10 (57.2%) (early vs. late genes: $P < 10^{-3}$ by chi-square test). C11 contained mRNAs whose abundance poorly correlated with time. As a negative control, a cluster C12 made with 50 mRNAs randomly taken from all those that did not exhibit any change in this study provided a flat curve.

As an external control for the above selection, one mRNA taken from every cluster C1 to C10 was randomly tested by q-RT-PCR. In all instances, the kinetics of

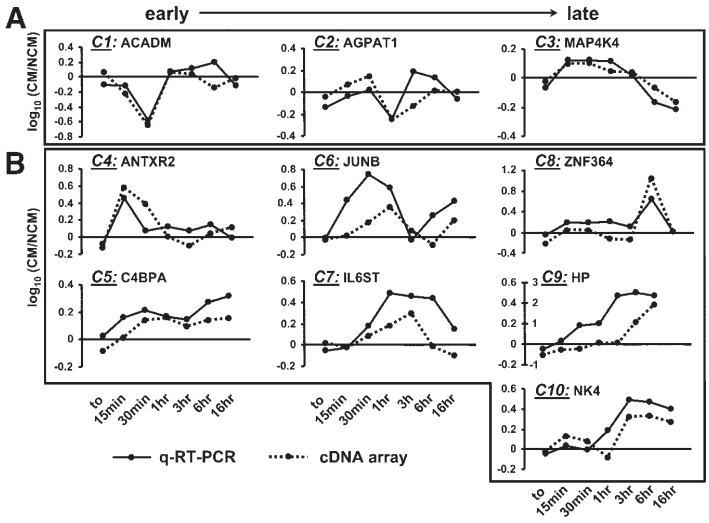


Fig. 2. Time course of abundance for selected mRNAs as controlled by q-RT-PCR. The clusters C1 to C10 and the general presentation are as in Fig. 1A. For every mRNA identified on top, the relative abundance in CM- versus NCM-challenged cells was determined in triplicate (microarray) or duplicate (q-RT-PCR) experiments and shown as a log ratio. In C9, the 2 scales used for microarray (left ordinate) or q-RT-PCR (right ordinate) are different. ACADM, acyl-coenzyme A dehydrogenase, C-4 to C-12 straight chain; AGPAT1, 1-acylglycerol-3-phosphate *O*-acyltransferase 1; MAP4K4, mitogen-activated protein kinase kinase kinase kinase 4; ANTXR2, anthrax toxin receptor 2; C4BPA, complement component 4 binding protein, alpha; JUNB, jun B proto-oncogene; IL6ST, membrane glycoprotein gp130; ZNF364, zinc finger protein 364; HP, haptoglobin; NK4, natural killer cell transcript 4; q-RT-PCR, quantitative reverse transcription polymerase chain reaction; CM, conditioned medium; NCM, nonconditioned medium.

abundance as found by microarray or q-RT-PCR were quite similar (Fig. 2). Moreover, for most of these mRNAs, the direction and kinetics of change in abundance were quite similar in CM-challenged Hep3B or HepG2 hepatoma cells (Supplementary Fig. 1), which makes a cell line–specific effect unlikely.

Abundance of Functionally Identified mRNA Subpopulations and Their Kinetics. We first verified that our series of 648 Hep3B/CM mRNAs fit a pro-inflammatory cytokine-induced regulation. In Fig. 3, many Hep3B/CM mRNAs that code for (1) critical proteins of the major cytokine-driven cascades or (2) other proteins in the hepatocyte under acute inflammation[2,6-8,11,12,21-26] were regulated as expected.

The mRNAs that code for proteins involved in (1) the immune response at large or (2) the inflammatory response were further identified by an ontology approach.[20] When comparing the numbers of such mRNAs found in either the Hep3B/CM mRNA population (see details in Supplementary Table 2) or in the entire population of mRNAs in the quiescent liver,[12] both functional groups were significantly enriched in the former (in both instances, $P < 10^{-4}$ by chi-square test). Again, this demonstrates that our selection of the Hep3B/CM mRNAs fits with a major influence of pro-inflammatory cytokines on mRNA abundance.

We searched for a time-dependent regulation of other, functionally defined mRNA subpopulations in CM-challenged Hep3B cells. By ontology, we identified 13 major subpopulations of mRNAs corresponding to proteins with a well-identified function (Supplementary Table 2). Among the clusters C1 to C11, 7 clusters contained at least 1 significantly under- or over-represented mRNA subpopulation (Supplementary Table 3). A further comprehensive analysis is presented in Fig. 1B and Supplementary Fig. 2. Striking observations include (1) a trend to transcriptional upregulation (boxes C4-C5 and C6-C7) and translational repression (C1) at the early-to-intermediate phase of the cell response, (2) a repression of detoxication and hepatic metabolism in the intermediate phase (C2), and (3) an upregulated synthesis and transport of membranous or secreted proteins as well as increased protein degradation in the late phase (box C8-C10).

Stability-Dependent mRNA Abundance. We examined to which extent the stability of our set of 648 Hep3B/CM mRNAs was affected after a CM-versus-NCM challenge for either 30 minutes or 16 hours. The abundance of every mRNA was determined at various times after transcription arrest by actinomycin D. Genome-wide determinations of RNA stability still await standardized analysis.[14] Therefore, our specific calculations are summarized in Supplementary Table 4. After 30 minutes of CM challenge, 2 Gaussian populations of mRNAs were observed, as shown in Fig. 4A, left. One population (mean slope value = 0) had a narrow variation of stability that was considered unchanged by CM, and this was used to calculate a normal range of values (horizontal thick bar in Fig. 4A). Within the other population, the mRNAs had highly variable slope values, which significantly departed ($P < .05$) from the normal range. Quite similar data were obtained after a CM challenge for 16 hours (Fig. 4A, right). Altogether, 218 Hep3B/CM mRNAs did not

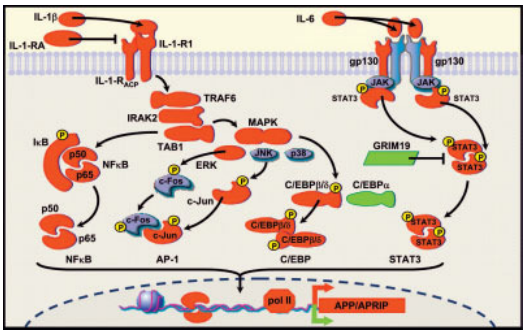


Fig. 3. Appropriate changes in mRNA abundances connected to proinflammatory cytokine-regulated pathways in CM-challenged Hep3B cells. The major proinflammatory cytokine-regulated pathways and relevant transcription factors are summarized: (1) the IL-1β/NF-κB pathway, (2) the IL-1β/MAP kinase/AP-1 and C/EBP pathway, and (3) the IL-6/STAT3 pathway. The abundance of an mRNA for a protein depicted in red or green was up- or downregulated in accordance with the literature. In particular, the AP-driven synthesis of STAT-3 and C/EBPβ and -δ molecules is a relatively late event *in vivo* and takes place only after latent STAT-3 and C/EBPβ molecules of the quiescent hepatocyte are activated by phosphorylation.[3,6] Hence, the upregulation of STAT-3 and C/EBPβ and -δ mRNAs (seen in C10) strongly suggests that our kinetics covers most of the early to intermediate stages of the hepatocyte response to pro-inflammatory cytokines. Also in keeping with this, the mitogen-activated protein kinase 1 and IκBα mRNAs that allow for a temporal control of NFκB activity in AP[22,23] were found to be upregulated (in C4 and C5, respectively). Some mRNAs for other proteins that were not found to be regulated are also indicated (in gray) as they fill major gaps in the summarized pathways. A critical phosphorylation event is indicated with a P in a yellow circle. The nucleus membrane is symbolized with a curved dotted line. The promoter of an APP- or APRIP-encoding gene is symbolized at the bottom (red or green broken arrow: transcription start site for an up- or downregulated gene). AP-1, activating protein-1 (c-Jun/c-Fos dimer); APP, acute phase protein; APRIP, acute phase-regulated intracellular protein; C/EBP, CCAAT-enhancer binding protein; ERK, extracellular signal-regulated kinase (=MAPK1); gp130, glycoprotein of 130 kd; GRIM-19, gene associated with retinoid-interferon-induced mortality 19; IκB, inhibitor of NF-κB; IL1-RA, IL-1β receptor antagonist; IL1-RACP, IL-1β receptor accessory protein; IL1-R1, IL-1β receptor type 1; IRAK2, IL-1β receptor-associated kinase 2; JAK, Janus kinase; JNK, c-Jun kinase; MAPK, mitogen-activated protein kinase; NF-κB, nuclear factor kappa B; p38, 50, 65, protein of 38, 50 or 65 kd; pol II, RNA polymerase II; STAT, signal transducer and activator of transcription; TAB1, transforming growth factor β-activated protein kinase 1-binding protein 1; TRAF6, Tumor necrosis factor receptor–associated factor 6. Some other inflammation-associated mRNAs (not shown) code for (1) inflammation-driven receptors and cytokines (*e.g.,* CD14 and TNFα, peaking in C6 and C4, respectively); (2) associated signal transduction (*e.g.,* the TNF receptor-associated factor-4, peaking in C4-C5); (3) transcription factors and related proteins (*e.g.,* HNF-3α and HNF-4α decreased in C2; the c-Jun activator MAP4K4, decreased in C3; ATF3, peaking in C5; JunB and STAT5a and -5b peaking in C6); (4) HLA-B, -C, and -class II chains (upregulated in C7 and C10), and the class I regulator tapasin (downregulated in C2); and (5) blood-borne APPs (*e.g.,* albumin and fetuin-A and -B decreased in C3; various complement components, C-reactive protein, haptoglobin, fibrinogen, and orosomucoid peaking in C9-C10). The direction and kinetics of all these changes fit many published data.[2,6-8,11,12,21-26]

present any variation of stability, whereas the other 430 Hep3B/CM mRNAs exhibited a CM-induced change of stability (the latter mRNAs are noted as such in Supplementary Table 2). As seen in Fig. 4A, the mRNAs with a decreased stability (negative slope) predominated (68.4% of all 430 mRNAs) as compared with those with an enhanced stability (31.6%) after 30 minutes as well as 16 hours of CM challenge. As shown in Fig. 4B, the proportion of mRNAs with an enhanced or decreased stability was not significantly different in box C1-C3 (decreased abundance) compared with box C4-C10 (increased abundance), thus ruling out that altered stability alone accounted for an up- or downregulated mRNA abundance. However, the relative distribution of mRNAs with altered stability in the clusters C1 to C10 was not random. The number of mRNAs with enhanced stability was significantly higher, and the number of unstable mRNAs was significantly lower, in box C8-C10 (clusters of increased abundance) as compared with C3 (decreased abundance) (stars in Fig. 4B). This stability/abundance relationship is logical and validates our experimental approach. Moreover, this relationship increased timewise from box C4C5 up to C8-C10, whereas it was not observed when comparing C1, C2, and C3. Taken together, our data indicate that (1) a loss of stability controls, at least partly, the abundance of many AP mRNAs and (2) stability enhancement occurs infrequently and mostly acts on the late mRNAs.

Because mRNA 3′ UTR may be involved in stability,[14] we investigated whether the stability-regulated mRNAs identified above exhibited some characteristic physical features. As shown in Supplementary Table 5, the 3′ UTR of the mRNAs with modified stability was significantly shorter than that of control mRNAs, which fits earlier observations made at the level of absolute decay rate.[14] Moreover, the frequency of occurrence of 3 AREs that are known to be associated with a change in mRNA stability[13,14] was significantly lower in regulated than in control mRNAs. These data support our identification of mRNA populations with a modified or unchanged stability post-CM challenge.

Transcriptional Control of mRNA Abundance. We also examined to what extent the variations in abundance resulted from a change in transcription in our set of 648 Hep3B/CM mRNAs after 30 minutes or 16 hours of CM challenge. Comparing the relative abundance of primary transcripts in nuclei from CM- versus NCM-challenged cells identified 191 or 66 transcripts that were significantly regulated at 30 minutes or 16 hours, with very little overlap (Fig. 5A). They are so noted in Supplementary Table 2. Remarkably, the downregulated primary tran-

scripts predominated at 30 minutes (143 of 191, 74.8%), but they were a minority at 16 hours (21 of 66, 31.8%) ($P < 10^{-4}$ by chi-square test). This indicates that transcriptional repression predominates during the early AP, whereas transcriptional activation predominates at a later stage. As shown in Fig. 5B, the fraction of primary transcripts indicative of a decreased or enhanced transcription were not significantly different in the mRNAs of decreased abundance (box C1-C3) versus those of increased abundance (box C4-C10), thus ruling out that transcription alone accounted for an up- or downregulated mRNA abundance. However, the number of upregulated primary transcripts appeared to be higher and the number of downregulated transcripts lower in box C8-C10 (mRNAs with a late increase in abundance) as compared with C3 (late decrease), although this was not significant because of the small sample size. This transcriptional activity/mRNA abundance relationship is logical and supports our run-on analysis. Taken together, our data indicate that (1) altered transcription controls, at least partly, the abundance of many AP mRNAs, and (2) increased transcription mostly affects the late AP genes.

Within the mRNAs whose transcription and stability both were found to be altered in this study, those with opposite regulations of transcription and stability were as numerous as those with 2 up- or downregulations ($P > .5$, not detailed), and the former were evenly distributed in all clusters. Therefore, additive or subtractive effects of transcription and stability have similar occurrence. Moreover, within the subset of mRNAs undergoing 2 up- or 2 downregulated transcription and stability, the extents of both parameters did correlate ($P < .05$, not detailed), suggesting cooperation. On the contrary, these parameters were anti-correlated within the subset of mRNAs with opposite transcription and stability ($P < .05$, not detailed), and hence transcription and stability may antagonize in an imbalanced fashion, which still results in unidirectional change in mRNA abundance. None of the functionally defined subpopulations previously noted in Fig. 1B and Supplementary Fig. 2 appeared to be preferentially associated with any transcription/stability combination (not detailed).

Cytokine-Induced Changes in the Polysome Fraction of mRNAs. Because changes in mRNA and protein levels do not necessarily correlate,[27] translation of a given

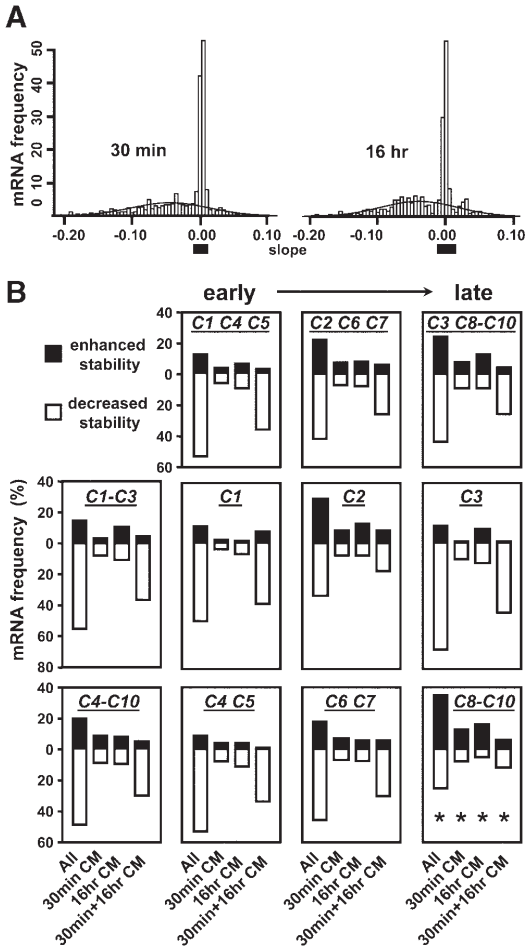


Fig. 4. CM-associated changes in mRNA stability. Transcription arrest by actinomycin D was done after 30 minutes or 16 hours of CM or NCM challenge. The abundance of every mRNA in the set of 648 Hep3B/CM mRNAs was measured by microarray at various times after actinomycin D (as detailed in Supplementary Table 4). (A) Distribution of values of mRNA stability. Abscissa, slope of the curve of relative mRNA abundance in the CM- versus NCM-treated cells at various times after actinomycin D. A negative (or positive) slope value indicates an mRNA with a CM-associated decrease (or enhancement) of stability. Ordinate: absolute number of mRNAs. At either 30 minutes or 16 hours of CM challenge, one mRNA population with an unchanged stability post-CM was distributed as a sharp, Gaussian peak that provided a normal range of values shown as an horizontal thick bar (mean ± 2 SD). Another population with a wide, Gaussian distribution (solid line) significantly departed from this normal range ($P < .05$). The total numbers of mRNAs with unchanged or altered stability were 218 and 430, respectively. At either 30 minutes or 16 hours after CM, the total numbers of mRNAs with enhanced/decreased stability were 74/237 and 96/245, respectively. (B) Distribution of the 430 mRNAs with a CM-altered stability in the clusters C1 to C10. Abscissa: The mRNAs were counted on the basis of an altered stability observed only after 30 minutes' (30min CM) or 16 hours' CM challenge (16hr CM) or after both durations (30min+16hr CM), or they were all counted together (All). Ordinate: within a box, the bars depict the number of mRNAs with an enhanced (solid bar) or decreased stability (open bar) expressed as a percentage of the total number of mRNAs in this box. In box C8-C10 versus box 3, a star indicates a statistically significant difference ($P < .03$ by chi-square test).

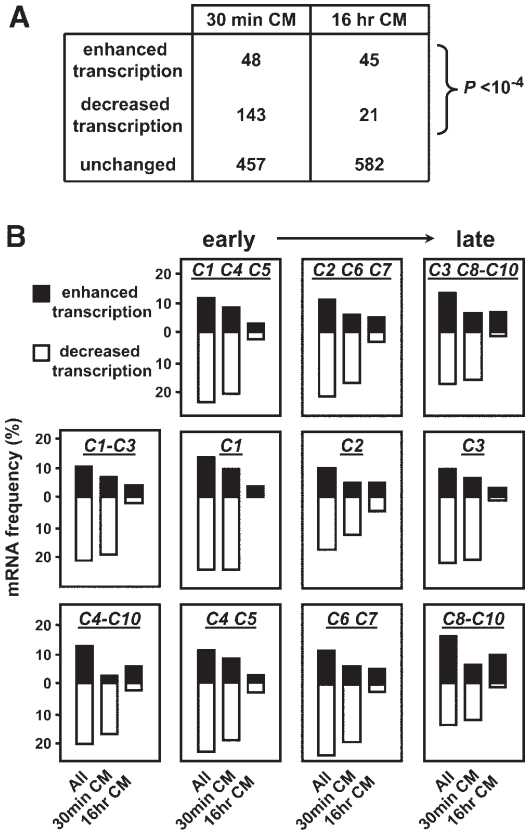


Fig. 5. CM-associated changes in primary transcripts. The abundance of primary transcripts was determined in the set of 648 Hep3B/CM mRNAs after 30 minutes or 16 hours of CM or NCM challenge by microarray. (A) Number of primary transcripts with a significantly altered abundance at either time. (B) Distribution of the transcripts with a CM-altered abundance in clusters. Abscissa: the mRNAs were counted on the basis of an altered abundance of the primary transcript observed only after 30 minutes' (30min CM) or 16 hours' CM challenge (16hr CM) or after they were all counted together (All). Ordinate: within a case, the bars depict the number of primary transcripts with an enhanced (solid bar) or decreased abundance (open bar) expressed as a percentage of the total number of mRNAs in this case.

mRNA could be CM-modulated, regardless of whether its abundance was altered. Therefore, and regardless of the 648 Hep3B/CM mRNAs listed, we searched for a CM-associated change in the ratio of (polysome-bound molecules/[free + monosome]-bound molecules) for every mRNA that was detectable in the Hep3B cells. As above, this was carried out at 30 minutes or 16 hours of CM challenge. The identification of mRNAs whose relative abundance in these 2 populations was most significantly ($P < .05$) modified by a CM challenge is illustrated in Supplementary Fig. 3. These mRNAs were considered to be engaged in a CM-dependent (up- or downregulated) change of translation extent. This resulted in a final selection of 34 or 40 such mRNAs at 30 minutes or 16 hours of CM challenge, respectively. At either time, approximately half of the mRNAs had undergone an increased translation, whereas translation of the remaining mRNAs was decreased (nonsignificant difference by chi-square test), thus indicating that the translational control of protein production in AP is bidirectional. The complete list of these mRNAs (Supplementary Table 6) shows very little overlap between the mRNAs whose translation is regulated at 30 minutes or 16 hours. In fact, no correlation existed between the extent of translation observed at 30 minutes and 16 hours for the 34 ($r = 0.002$, $P = .99$) or 40 mRNAs identified previously ($r = 0.13$, $P = .41$). This observation points to a strongly time-dependent control of translation for most AP-relevant proteins. Remarkably, the mRNAs with an altered translation included (1) actors of the inflammatory response, at 30 minutes (RAB family members) or 16 hours (*e.g.,* metallothionein 1H, leukemia inhibitory factor, MAP kinase kinase-1, TNF receptor superfamily member 11a), (2) prominent actors of protein degradation (ubiquitin protein ligase E3A and proteasome subunit β6, upregulated at 30 minutes), and (3) actors of translation (ribosomal proteins L41 and S23) whose upregulation at 16 hours suggested a positive feedback loop for a late enhancement of translation.

As noted in Supplementary Table 6 (last column), 8 mRNAs regulated in translation were previously found to be regulated in abundance, and with only one exception (cytokine-like nuclear factor n-pac) both regulation levels acted in the same direction. Strikingly, the 2 mRNAs with the most tightly up- or downregulated translation at 30 minutes of CM challenge were also regulated in abundance (RNA polymerase II polypeptide F; RAB18). Likewise, several mRNAs with a highly upregulated translation at 16 hours of CM challenge were also upregulated in abundance. In these situations, mRNAs belonging to the late clusters of upregulated abundance (C9, C10) predominated. Taken together, our data suggest that in a limited number of cases, shifts in mRNA abundance and translation can cooperate for an upregulated protein synthesis during the late AP, notably when a strong translational control takes place.

Protein electrophoresis and immunodetection were used as a control for translationally regulated mRNAs. Three mRNAs with a time- and CM-dependent shift in translation, namely, POLR-2F, CAM-1, and MAO-B, were selected. As shown in Supplementary Fig. 4, the abundance of the PolR-2F protein increased after a 30-minute CM challenge, whereas that of CAM-1 or

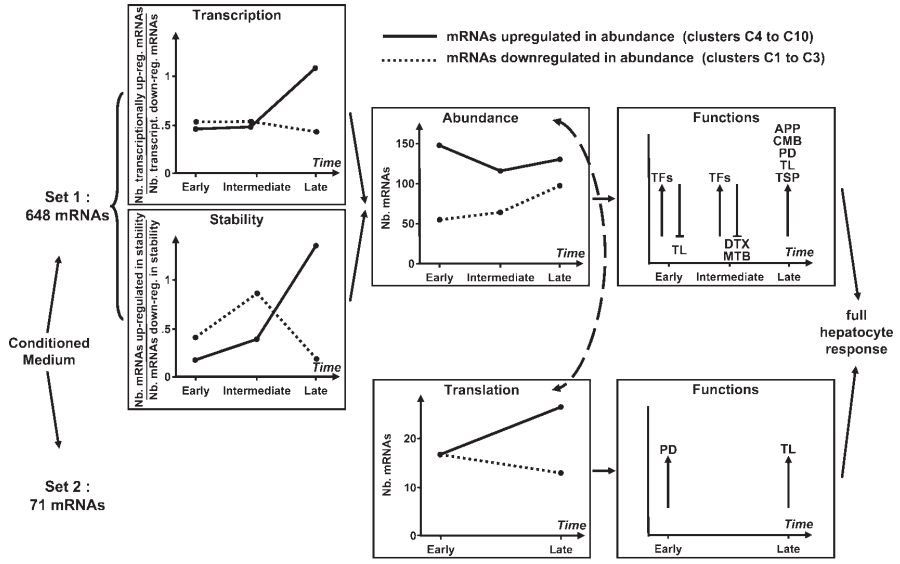


Fig. 6. Overall view of the CM-induced Hep3B cell response. After a pro-inflammatory cytokine challenge (Conditioned Medium), a set 1 made of 648 mRNAs is regulated in abundance, whereas another, essentially independent set 2 of 71 mRNAs is mainly or exclusively regulated translationally. In every panel, a control pathway or functional targets are noted at the top. An extreme case of combined control involves mRNA abundance and translation (double-headed, curved, and dotted arrow), which mostly applies to some mRNAs with a strong and late upregulation. Up- or downregulation of functionally identified protein group(s) is respectively depicted by an ascending arrow below- or an upside down T above the group name (TF, transcription factor; TL, translational machinery; DTX, detoxication; MTB, liver metabolism; APP, acute phase protein; CMB, cell membrane; PD, protein degradation; TSP, transport). The plotted values (closed circles in the 4 leftmost panels) were calculated from data found in Supplementary Table 2 (for the transcription and stability panels), Fig. 1A (abundance panel), and Supplementary Table 6 (translation panel).The "functions" panels summarize data taken from Fig. 1B and Supplementary Fig. 2 and Supplementary Tables 3 and 6.

MAO-B decreased after a 30-minute or 16-hour CM challenge, respectively. These data cannot be accounted for by a concomitant change in mRNA abundance (lower diagrams in Supplementary Fig. 4), and hence they support our identification of the mRNAs with a CM-dependent translation.

Discussion

The percentage of regulated mRNAs found in this study (approximately 7%) is quite similar to the number of liver mRNAs regulated during the AP in human or mouse *in vivo*.[11,12] Most importantly, the proper direction and kinetics of changes in mRNA abundances for cytokines and their receptors, transcription factors (TFs), and APPs all support our choice of the Hep3B cells. Our further data (not shown) did not indicate any significant modulation of mRNAs for the suppressors of IL-6 signaling that are involved in the resolution phase of inflammation.[3,4] This is consistent with a sustained CM stimulation along the time course of this study, which likely prevented the target Hep3B cells from returning to a quiescent state. Moreover, the discrete waves of stimulation by TNFα and IL-1β, then IL-6, which occur *in vivo*,[1,2] could hardly be mimicked timewise *in vitro*, which may have prevented the resolution phase from occurring. Finally, we preferred not to study the effect of a single proinflammatory cytokine. For instance, IL-1β is known to promote an opposite effect on some APPs, depending on whether it acts in the context of other cytokines such as IL-6.[1,2,6]

The over- or under-representation of functional groups along the AP time course discloses an early control of TF-encoding mRNAs, which subsequently results in an up- or downregulated production of other mRNAs for proteins with mostly hepatocyte-specific functions. Downregulation of enzymes involved in detoxication and metabolism (mostly in C2) represents a noticeable example, which includes several alcohol dehydrogenases and glutathione transferases, as well as key enzymes for metabolism of glucose (glucose-6-phosphatase), fatty acids (stearoyl-CoA desaturase) or cholesterol (24-dehydrocholesterol reductase). This downregulation can be at least partly accounted for by the concomitant downregulation of HNF-4 and upregulation of STAT-3, because the

former is an activator of hepatocyte-specific metabolism at large,[28] and the latter is a repressor of the glucose-6-phosphatase–dependent gluconeogenesis.[29] This transient downregulation takes place in a highly time-sensitive manner, as it disappeared during the late phase of cytokine challenge in this study, despite the possible lack of a resolution phase as discussed. An enhanced production of some APPs in the AP could benefit from amino acid saving resulting from a decreased synthesis of other proteins.[2] We now demonstrate that the transient downregulation of detoxication and metabolism takes place before the increased synthesis of a bulk of APPs, which argues for a participation of some transiently dispensable detoxication and metabolism proteins in such an amino acid–saving scenario.

Contrary to other reports in which stability was merely inferred from combined transcriptional rate and mRNA abundance,[30,31] we actually measured the CM-associated change of mRNA decay. We have now found that two thirds of the Hep3B/CM mRNAs are regulated by stability. This is in keeping with the change of abundance noticed for mRNAs coding for several heterogeneous nuclear ribonucleoproteins (hnRNP), and particularly hnRNP D (cluster C5), which directly controls mRNA stability.[32] Predominance of mRNAs with a loss of stability is a further novel finding of our study. It should not be seen as a standard response to stress, given that a shift toward stabilization has been found in other examples of cellular stress.[30,31] The current loss of stability likely results, at least in part, from the early repression of the MAP kinase-2 mRNA (cluster C1) that limits mRNA decay.[33] It also appears to be driven by AREs, given the lower frequency of AREs found in the 3′UTR of regulated mRNAs versus controls in our study. AREs participate in decay control but they do not allow prediction of the direction and extent thereof.[14,33] Finally, the predominant loss of stability seen in our AP-regulated mRNAs is consistent with a requirement for (1) a transient downregulation of some mRNAs controlling normal functions in the quiescent hepatocyte (*e.g.,* metabolism) and (2) a short-term limitation of some AP-induced mRNAs whose sustained presence could be detrimental.

Analysis of nuclear transcripts by run-on has seldom been made on a genome-wide scale.[30] We developed this approach in the hepatocyte/AP context and, as expected,[6,9] we observed that transcription controls a number of APRIP/APP genes. The overall trend of this control step is time-dependent, as transcriptional repression or activation predominates at an early or late stage of AP, respectively. This feature has not previously been documented. It fits the up- or downregulated abundance of many TF-encoding mRNAs, and kinetics thereof, within our set of Hep3B/CM mRNAs. Not only can such an early decrease of given TFs directly account for a subsequent limitation of other proteins (*e.g.,* the direct relationship between HNF-4 and metabolism-related proteins) but it also can allow for a subsequent upregulation of other TFs. For instance, GRIM-19, a STAT-3 inhibitor whose mRNA decreases early, allows for (1) an immediate upregulation of STAT-3 activity and (2) a late upregulation of STAT-3 targets, such as the *C/EBPβ* gene.[6]

Within every cluster C1 to C10, some mRNAs exhibited a change in either stability or transcription that was opposed to their final change in abundance. This feature fits with other reports of opposite regulations of mRNA transcription, stability, or translation in various contexts.[14,30,31] This now underscores that in the AP-challenged hepatocyte transcription and stability can either cooperate or antagonize and still result in an unidirectional change of abundance. Potentially conflicting data appeared in the early phase of our kinetics, as we found (1) a predominance of mRNAs with an increased abundance, along with (2) predominant numbers of transcriptionally repressed mRNAs and mRNAs with a decreased stability. The fact that two counteracting regulations of transcription and stability often occur and still result in a unidirectional change of abundance clarifies, at least partly, this conflict.

This study provides a genome-wide and kinetic view of the human hepatocyte response to pro-inflammatory cytokines, from transcription to translation. Because the data obtained by our various approaches cover quite different scales, trends rather than absolute figures should be compared. Such trends shown in Fig. 6 (left 2 panels) include an overall predominance of mRNAs with a CM-induced decrease in stability, and an increased transcription and stability of the mRNAs whose abundance is upregulated late in the course, along with reversed regulations of the mRNAs with a downregulated abundance. Among the latter, those coding for elements of the translational machinery are repressed early, whereas translation is re-activated later (right 2 panels), which fits with a timewise increase in the number of mRNAs that are unaltered in abundance but translationally upregulated (center 2 panels). Therefore, translation represents a critical control for APRIP/APP production. In extreme cases of upregulation, mRNA abundance and translation cooperate (curved, dotted arrow between panels). Our overall view fits with: (1) an engagement of latent proteins and translation of a limited number of latent mRNAs as the primary events at the onset of the hepatocyte response, the latter also including a repressed translation or active

degradation of other pre-existing proteins (*e.g.,* IκB) and (2) a regulation of gene activity and protein synthesis that reaches its full extent at a later stage and allows for the full cell response to take place. The latter notably includes a repression of detoxication and metabolism, an enhanced translation and transport of a number of membranous or secreted proteins, as well as an enhanced protein degradation, which may in turn limit the production of upregulated but potentially harmful proteins[34] in a time-dependent manner.

References

1. Baumann H, Gauldie J. The acute phase response. Immunol Today 1994; 15:74-80.
2. Gabay C, Kushner I. Acute-phase proteins and other systemic responses to inflammation. N Engl J Med 1999;340:448-454.
3. Heinrich PC, Behrmann I, Haan S, Hermanns HM, Muller-Newen G, Schaper F. Principles of interleukin (IL)-6-type cytokine signalling and its regulation. Biochem J 2003;374:1-20.
4. Wormald S, and Hilton DJ. Inhibitors of cytokine signal transduction. J Biol Chem 2004;279:821-824.
5. Olivier E, Soury E, Risler JL, Smih F, Schneider K, Lochner K, et al. A novel set of hepatic mRNAs preferentially expressed during an acute inflammation in rat represents mostly intracellular proteins. Genomics 1999; 57:352-364.
6. Ruminy P, Gangneux C, Claeyssens S, Scotte M, Daveau M, Salier JP. Gene transcription in hepatocyte during the acute phase of a systemic inflammation: from transcription factors to target genes. Inflamm Res 2001;50:383-390.
7. Soury E, Olivier E, Simon D, Ruminy P, Kitada K, Hiron M, et al. Chromosomal assignments of mammalian genes with an acute inflammation-regulated expression in liver. Immunogenetics 2001;53:634-642.
8. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Prot 2003;1:845-867.
9. Jiang SL, Samols D, Rzewnicki D, Macintyre SS, Greber I, Sipe J, et al. Kinetic modeling and mathematical analysis indicate that acute phase gene expression in Hep3B cells is regulated by both transcriptional and post-transcriptional mechanisms. J Clin Invest 1995;95:1253-1261.
10. Chinnaiyan AM, Huber-Lang M, Kumar-Sinha C, Barrette TR, Shankar-Sinha S, Sarma VJ, et al. Molecular signatures of sepsis: multiorgan gene expression profiles of systemic inflammation. Am J Pathol 2001;159:1199-1209.
11. Yoo JY, Desiderio S. Innate and acquired immunity intersect in a global view of the acute-phase response. Proc Natl Acad Sci U S A 2003;100: 1157-1162.
12. Coulouarn C, Lefebvre G, Derambure C, Lequerre T, Scotte M, Francois A, et al. Altered gene expression in acute, systemic inflammation detected by complete coverage of the human liver transcriptome. HEPATOLOGY 2004;39:353-364.
13. Bakheet T, Williams BR, Khabar KS. ARED 2.0: an update of AU-rich element mRNA database. Nucleic Acids Res 2003;31:421-423.
14. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res 2003;13:1863-1872.
15. Daveau M, Jean L, Soury E, Olivier E, Masson S, Lyoumi S, et al. Hepatic and extra-hepatic transcription of Inter-α-Inhibitor family genes under normal or acute inflammatory conditions in rat. Arch Biochem Biophys 1998;350:315-323.
16. Zong Q, Schummer M, Hood L, Morris DR. Messenger RNA translation state: the second dimension of high-throughput expression screening. Proc Natl Acad Sci U S A 1999;96:10632-10636.
17. Mikulits W, Pradet-Balade B, Habermann B, Beug H, Garcia-Sanz JA, Mullner EW. Isolation of translationally controlled mRNAs by differential screening. FASEB J 2000;14:1641-1652.
18. Arava Y. Isolation of polysomal RNA for microarray analysis. Methods Mol Biol 2003;224:79-87.
19. Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. Bioinformatics 2002;18:207-208.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25-29.
21. Gangneux C, Daveau M, Hiron M, Derambure C, Papaconstantinou J, Salier JP. The inflammation-induced down-regulation of Fetuin-A (α2HS-Glycoprotein) in liver results from the loss of interaction between long C/EBP isoforms at two neighbouring binding sites. Nucl Acids Res 2003;31:5957-5970.
22. Baeuerle PA. IkappaB-NF-kappaB structures : at the interface of inflammation control. Cell 1998;95:729-731.
23. Jiang B, Xu S, Hou X, Pimentel DR, Brecher P, Cohen RA. Temporal control of B activation by ERK differentially regulates interleukin-1β-induced gene expression. J Biol Chem 2004;279:1323-1329.
24. Ripperger JA, Fritz S, Richter K, Hocke GM, Lottspeich F, Fey GH. Transcription factors Stat3 and Stat5b are present in rat liver nuclei late in an acute phase response and bind interleukin-6 response elements. J Biol Chem 1995;270:29998-30006.
25. Zhang J, Yang J, Roy SK, Tininini S, Hu J, Bromberg JF, et al. The cell death regulator GRIM-19 is an inhibitor of signal transducer and activator of transcription 3. Proc Natl Acad Sci U S A 2003;100:9342-9347.
26. Wright CA, Kozik P, Zacharias M, Springer S. Tapasin and other chaperones: models of the MHC class I loading complex. Biol Chem 2004;385: 763-778.
27. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 2003;4:117.
28. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, et al. Control of pancreas and liver gene expression by HNF transcription factors. Science 2004;303:1378-1381.
29. Inoue H, Ogawa W, Ozaki M, Haga S, Matsumoto M, Furukawa K, et al. Role of STAT-3 in regulation of hepatic gluconeogenic genes and carbohydrate metabolism in vivo. Nat Med 2004;10:168-174.
30. Garcia-Martinez J, Aranda A, Perez-Ortin JE. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol Cell 2004;15:303-313.
31. Kawai T, Fan J, Mazan-Mamczarz K, Gorospe M. Global mRNA stabilization preferentially linked to translational repression during the endoplasmic reticulum stress response. Mol Cell Biol 2004;24:6773-6787.
32. Laroia G, Sarkar B, Schneider RJ. Ubiquitin-dependent mechanism regulates rapid turnover of AU-rich cytokine mRNAs. Proc Natl Acad Sci U S A 2002;99:1842-1846.
33. Winzen R, Gowrishankar G, Bollig F, Redich N, Resch K, Holtmann H. Distinct domains of AU-rich elements exert different functions in mRNA destabilization and stabilization by p38 mitogen-activated protein kinase or HuR. Mol Cell Biol 2004;24:4835-4847.
34. Recinos A III, Carr BK, Bartos DB, Boldogh I, Carmical JR, Belalcazar LM, et al. Liver gene expression associated with diet and lesion development in atherosclerosis-prone mice: induction of components of alternative complement pathway. Physiol Genomics 2004;19:131-142.

Péroraison

Chapitre 6

Discussion

6.1 Liverpool

Les choix concernant la conception de Liverpool furent très tôt orientés vers des technologies bien connues et maîtrisées :

- la PCR et le dépôt de ses produits sur la puce ;
- le marquage des produits de PCR par radioactivité.

En effet celles-ci sont d'abord d'un coût moins élévé que celui des autres technologies comme les oligonucléotides et la fluorescence. Elle sont par ailleurs utilisées avec succès dans d'autres travaux [230, 231] et la radioactivité à cette propriété de posséder à la fois une importante sensibilité et une émission linéaire du signal [232].

Ces choix ont permis une rapide maîtrise de la technologie si bien qu'aujourd'hui encore notre laboratoire est un lieu unique en Haute-Normandie doté d'un savoir-faire hautement spécialisé dans l'élaboration et l'analyse de puces à ADN radioactives.

Constituée d'approximativement 10000 sondes, Liverpool est une puce à ADN qui couvre l'ensemble du transcriptome hépatique humain comme aucun outil ne le proposait jusqu'alors. Afin de définir quelles allaient être ces sondes, nous avons développé un outil informatique capable de retrouver les ADNc représentants les gènes dont l'expression est au moins hépatique. Présenté sous la forme d'un programme écrit en PERL, cet outil permet au sein des fichiers *Hs.data* maintenus par le NCBI [1], de rechercher les séquences d'ADNc répondant aux critères de sélection suivants :

- nous nous sommes assurés que la séquence de la sonde est la séquence complémentaire de la séquence d'ARNm située le plus en position 5' de celui-ci. Cette contrainte nous permet d'augmenter la spécificité de chacun des clones afin de limiter les hybridations croisées ;
- nous nous sommes également asssurés de la présence de l'annotation des tissus d'expression pour confirmer l'expression hépatique des clones.

Nous avons enfin parfait la liste des sondes par des recherches d'une part bibliographiques et d'autre part au sein des banques d'ADNc spécialisées. Les résultats de l'utilisation de Liverpool avec des échantillons de tissus non-hépatiques puis d'autre part avec des hépatocytes isolés montrent clairement la spécificité hépatique de notre puce à ADN.

Enfin pour s'assurer de la précision de notre outil, nous avons appliqué une politique rigoureuse du contrôle de la qualité. Nous avons ainsi d'abord validé la séquence et donc l'annotation de 686 sondes sélectionnées aléatoirement (10% de sondes présentes sur la puce) ce qui nous a permi de relever un taux d'erreur faible en accord avec celui annoncé par le consortium IMAGE fournisseur des

[1] National Center for Biotechnology Information, Bethesda, États-Unis

clones utilisés comme sondes. Puis nous avons verifié la faible variation entre les duplicats de sondes repartis aléatoirement sur la puce afin de mettre en évidence l'abcence d'effets spatiaux. Nous avons enfin vérifié que la position relative des sondes sur l'ARNm a une faible influence sur le signal émis.
Nous démontrons de cette manière que Liverpool est un outil précis, sensible et spécifique dédié à l'étude du transcriptome hépatique humain.

6.2 LiverTools

Parallèlement à notre puce à ADN, nous avons créé au sein du laboratoire un outil de gestion et d'analyse des données associées à Liverpool et baptisé LiverTools. La structure de LiverTools lui confère des qualités indispensables.

6.2.1 Structure

LiverTools peut être décrit en utilisant un diagramme de déploiement (cf. page 40) lequel définit une représentation de LiverTools en quatre noeuds (cf. figure 6.1) : « ordinateur client », « serveur ABISS[2]-CRIHAN[3] », « serveurs externes » et « serveur Inserm U519 ». Chacun de ces noeuds a une fonction dédiée et interagit avec ses noeuds voisins par l'intermédiaire d'associations.

Par exemple le noeud « serveur Inserm U519 » autour duquel s'articule l'ensemble des autres noeuds, est celui qui coordonne toutes les communications inter-noeuds. En effet qu'il s'agisse de l'accès aux données, de leur actualisation ou de leur analyse, ces tâches sont toutes à l'initiative des serveurs du noeud « serveur Inserm U519 ».
L'éclatement de la structure en plusieurs noeuds accorde à chacun de ceux-ci une certaine autonomie et permet une évolution du sytéme en limitant les repercussions de celle-ci sur les noeuds voisins.

6.2.2 Accès aux données en réseau

LiverTools est un ensemble d'outils intégrés organisés dans un réseau, lequel suit un schéma classique en trois strates (cf. page 47). L'organisation en réseau permet d'utiliser différents protocoles de transfert des données adaptés aux tâches demandées. Ainsi le protocole HTTP[6] utilisé pour accéder à LiverTools permet l'utilisation de logiciels standards et de se connecter au serveur d'applications

[2]Atelier de Biologie Informatique Statistiques et Sociolinguistique
[3]Centre de Ressources Informatiques de Haute-Normandie.
[6]HyperText Transfer Protocole

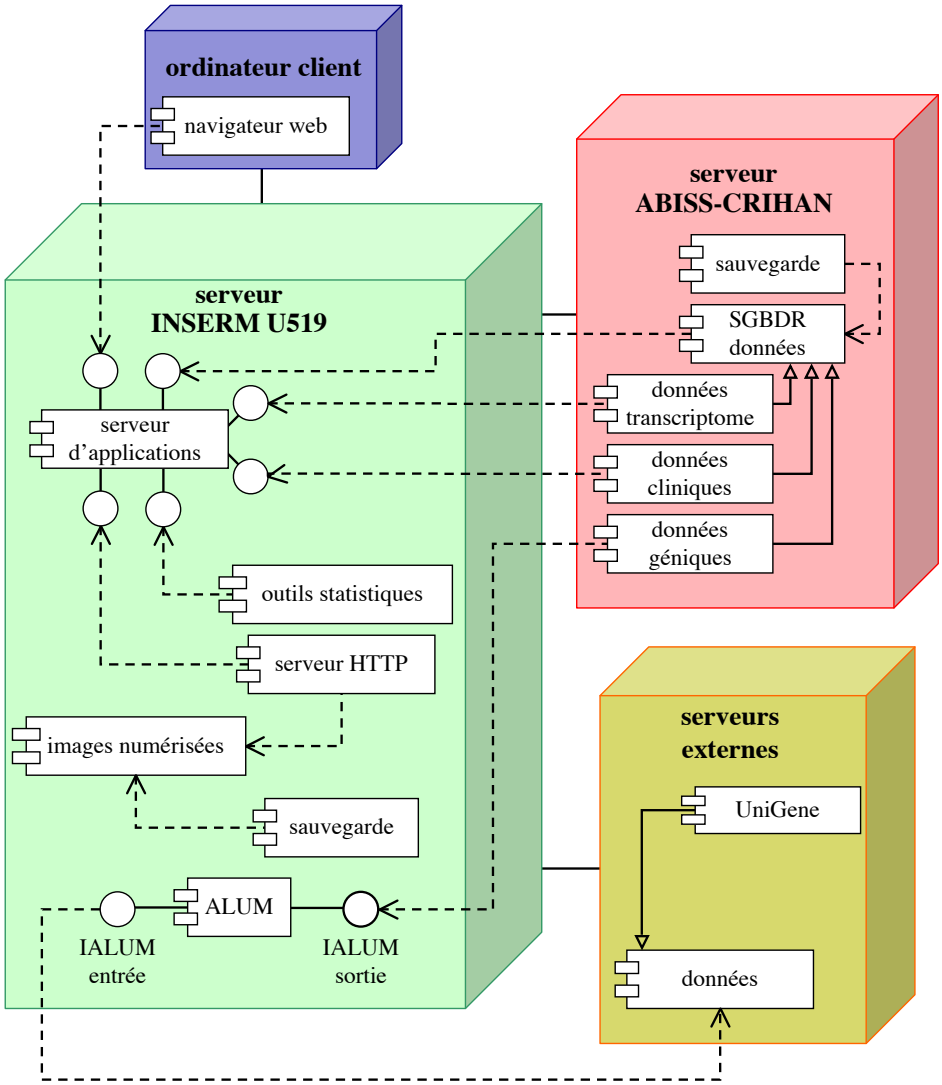


FIG. 6.1 – *Diagramme de déploiement de LiverTools.* LiverTools est constituée de quatre noeuds. Le premier est la voie par laquelle l'utilisateur accède à LiverTools grâce à un ordinateur « ordinateur client » et son navigateur internet. Le second noeud représente le « serveur ABISS-CRIHAN ». Il s'agit de l'ordinateur de ABISS dans les locaux du CRIHAN qui héberge le SGBDR. Son but est de maintenir les données et de permettre leur interrogation. Le troisième noeud est celui des « serveurs externes », sources d'informations pour maintenir actualisées les annotations des sondes. Il représente donc les conservatoires publiques tel le serveur d'UniGene hébergé au NCBI. Le quatrième enfin est le noeud « serveur Inserm U519 » qui représente l'ordinateur placé dans les locaux de l'unité 519 sur lequel sont basé plusieurs serveurs et modules d'intégration des données dont notamment le serveur d'application et le module ALUM[5]

depuis tous les postes informatiques du laboratoire. Par ailleurs LiverTools utilise le protocole NFS[7] ce qui permet de partager certains repertoires dont notament ceux contenant les images numérisées de la puce à ADN. Ces images peuvent alors être analysées depuis plusieurs postes informatiques par l'intermédiaire d'un logiciel d'analyse externe indépendament de LiverTools [159]. Ainsi au contraire d'un système hébergé par un seul poste informatique, LiverTools répond à un besoin de disponibilité accrue des données pour l'ensemble des protagonistes des projets utilisant Liverpool-LiverTools.

6.2.3 Pérennité et emmagasinnage des données

L'ensemble des données issues des expériences menées avec Liverpool est sauvegardé grâce à des systèmes indépendants s'appliquant chacun à un type de données particulier.
D'abord pour les images numérisées, j'ai déterminé une politique de sauvegarde qui utilise le système RAID[8] de niveau 1, grâce auquel chacune des images est dupliquée simultanément sur un disque dur différent. Ainsi si un disque dur sur lequel les images sont écrites est en panne, celles-ci ne sont pour autant pas perdues puisqu'elles restent accessibles sur le second disque dur. Par ailleurs une copie de chaque image est également effectuée sur un support amovible externe. Concernant ensuite les données textuelles, *i.e.* les données de transcriptome, cliniques et les annotations des sondes, celles-ci sont sauvegardées au bon soin du CRIHAN sur un système à bandes magnétiques. LiverTools suit donc une politique de sauvegarde des données qui assure la périnnité des données issues des expériences menées sur Liverpool.

Le respect des recommandations MIAME lors de l'emmagasinage des données (cf. page 55) est à mon sens aujourd'hui indispensable à plusieurs titres. Le premier est d'abord pragmatique. En effet cetaines revues dont les plus prestigieuses imposent de soumettre les paramètres expérimentaux aux bases de données officielles afin de se voir attribuer un numéro d'identification qui est alors demandé par leurs critiques pour qu'ils puissent étudier ces paramètres. Le second est plus conceptuel et est une conséquence d'une caractéristique majeure des puces à ADN aujourdh'hui et qui se traduit en un mot : « multiplicité ». Multiplicité des plateformes et des protocoles disponibles (Affymetrix™, Nimblegen™, Illumina™et nombre d'autres développées singulièrement par les laboratoires), multiplicité des termes d'annotation des sondes déposées sur les puces mais également multiplicité des méthodes d'analyses. Une telle diversité des paramètres impose, afin de permettre la confrontation des résultats des nombreuses études menées sur puces

[7]Network File System
[8]Redundant Array of Inexpensive Disks

à ADN, l'enregistrement de l'ensemble des paramètres expérimentaux [233].
C'est la raison pour laquelle nous avons souhaité dès la conception de LiverTools que le schéma de la base de données de LiverTools soit conforme aux principales recommandations MIAME.

6.2.4 Actualisation des annotations

Le *module automatisé d'actualisation de LiverTools* (Automated LiverTools Update Module) est un programme écrit en PERL intégré à LiverTools et dont le premier rôle est de vérifier périodiquement le niveau d'actualisation appelé également *construction*, des données contenues dans les serveurs d'UniGene du NCBI. Lorsque c'est utile, ALUM entre alors dans son second rôle et synchronise les données d'UniGene et celles de LiverTools. L'ensemble des fichiers UniGene concernant l'humain sont téléchargés puis parcouru et leurs informations sont enfin comparées à celles de la base de données de LiverTools. Lorsque l'ensemble des modifications à apporter est repertorié, ALUM envoi un ensemble de requêtes SQL au SGBDR afin d'actualiser les données contenues dans la base de données de LiverTools.
Bien que le facteur limitant essentiel soit le temps de téléchargement des fichiers, j'ai préféré cette méthode à celle utilisant des liens hypertextes inclus dans les pages HTML car cette dernière rend LiverTools constament dépendante d'une part de la validité de liens hypertextes qui sont très souvent éphémères et d'autre part de l'accessibilité des serveurs qui peut parfois être interrompue.

6.2.5 Mesure de la différence d'expression

Outre les outil de transformations des données brutes, LiverTools est dotée d'un outil qui assure la détermination des gènes dont l'expression est altérée par le phénomène biologique étudié. Pour cela j'ai traduit en R [234] l'algorithme proposé par Novak [193] et ses collaborateurs. Fondé sur l'observation que la dispersion des valeurs d'intensités est bien approximée par une fonction linéaire des valeurs d'intensité, le programme se décompose en trois phases :

- la première organise en ordre croissant chacune des valeurs en fonction de son intensité moyenne (μ) puis calcule la différence de ses intensités (Δ) entre les deux expériences comparées ;
- la seconde crée des groupes de n gènes consécutifs et calcule la dispersion des différences $\sigma(\Delta)_n$;
- la troisième détermine les paramètres de la fonction linéaires $\sigma(\Delta) = f(\mu)$, lesquels permettent de déterminer les équation des deux courbes limites au delà desquelles les différences d'intensité s'écartent signicativement des

valeurs dites normales et peuvent alors être attribuées avec un risque *alpha* d'erreur au phénomène biologique étudiés par l'expérience.

La simplicité de la méthode au regard de la qualité des résultats a été un atoût majeur en faveur de sa sélection.

Cependant je souhaiterais ici relever deux caractéristiques importantes.

La première est la limite du schéma expérimental accepté. En effet les échantillons ne sont comparables que deux à deux ce qui oblige à l'utilisation de mélanges, qu'ils soient réels dans une éprouvette ou virtuels par l'intermédiaire de valeurs moyennes.

La seconde est la propriété canonique de cette méthode qui toujours déterminera une proportion de gènes différentiellement exprimés correspondant au risque α d'erreur que l'expérimentateur s'autorise. L'archétype de cette propriété est la détermination de gènes dont l'expression est significativement altérée alors que les échantillons étudiés sont des réplicats. Certes mais dans ces conditions, les seules sources de variations sont d'origines techniques ou d'origines phénotypiques (variation individuelle). Or les variations induites par un phénomène biologique sont probablement bien plus importantes que celles induites par les deux sources précédemment citées. Et lors de la comparaison d'échantilllons placés dans des états physio-pathologiques différents, les différences d'intensités significativement anormales sont très probablement, au risque α accepté, dûes majoritairement au phénomène biologique objet de l'étude plus qu'à la propriété précédemment citée de la méthode.

Nous avons ainsi créé au sein de l'unité Inserm 519 un outil qui nous a permis de mettre en lumière des acteurs et des phenomènes essentiels intervenant durant la phase aiguë de l'inflammation systémique.

6.3 Marqueurs plasmatiques potentiels

Un résultat important des travaux présentés est la découverte de nouveaux marqueurs de la phase aiguë. Afin de déterminer la pertinence de ces marqueurs, nous avons défini un score de la phase aiguë fondé d'une part sur huit paramètres clinico-biologiques et d'autre part sur la classification en deux groupes de patients ; l'un constitué des malades et l'autres des témmoins.

Comme attendu, nous avons découvert parmi l'ensemble des gènes exprimés durant la phase aiguë, des gènes, au nombre de 154 dont 134 nouvellement associés à la phase aiguë. Leur niveau d'expression est proportionel au score de phase aiguë préalablement défini, ce qui est notament remarquable pour les APP positives plasmatiques. Par ailleurs le regroupement des patients basés sur le niveau d'expression de ces gènes permet de retrouver la classification initialement

établie sur critères clinico-biologiques.
Par ailleurs, les échantillons étant des prélèvements de tissus péri-tumoraux, d'aucuns pourraient arguer de l'origine non hépatocytaire des marqueurs nouvellement découverts. Ainsi avons-nous montré par Q-RT-PCR sur la lignée hépatome Hep3B stimulée *in vitro* par des cytokines pro-inflammatoires que l'abondance relative de ces cinq marqueurs était identique, à l'exception de TRAF5. Il est très probable que cette exception soit en fait une imprécision dûe au grand écart de temps entre les mesures (30min,16hrs).
Nous avons enfin vérifié par un test non-paramétrique, la présence extraordinairement élevée dans les séquences promotrices de ces gènes, de l'ensemble des motifs de fixation des principaux facteurs de transcription associés à la phase aiguë de l'inflammation.
Pouvons-nous ainsi raisonnablement définir ces 154 gènes comme marqueurs potentiels de la phase aiguë de l'inflammation ? Nous répondons assurément par l'affirmative à cette question. En effet qu'elle soit positive ou négative, la corrélation de leur niveau d'expression avec le degré de l'inflammation permet de définir les modulations de leur expression comme une marque potentiellement fiable de la phase aiguë de l'inflammation. Nous relèverons enfin que parmi les 154 gènes mis en lumière lors de ce travail, 5 nomément le *natural killer cell transcript 4* [Nk4], la *insulin-like growth factor binding protein 2* [IG-FBP2], *TGF-β-3*, la *laminine-α-5* ainsi que le *collagen-IV-α-1* codent pour des protéines très probablement plasmatiques [235], les plaçant alors dans la tête de liste des marqueurs cliniques potentiels de la phase aiguë de l'inflammation et rejoignant ainsi les marqueurs déja utilisés par les médecins cliniciens tels la CRP (*C-Reactive Protein*), l'haptoglobine, l'orosomucoïde, l'albumine et la transferrine.

6.4 Stimulus cytokinique et conséquences

6.4.1 Choix du modèle et du stimulus

Afin de maintenir un état de phase aiguë de l'inflammation, il était important de contrôler finement les paramètres expérimentaux et particulièrement au niveau même de l'hépatocyte.
Or l'entrée en phase de résolution de l'inflammation est précédée de vagues de régulations des cytokines. Parmi ces cytokines, le TNFα, l'IL-1β puis notamment l'IL-6 sont les principaux protagonistes de ce phénomène [12, 20]. Cette contrainte inscrite dans notre démarche expérimentale impose alors l'utilisation d'un modèle *in vitro*. L'inhibition de ces régulations était alors une clef du maintien en phase aiguë de l'inflammation et c'est pour cette raison que notre démarche expérimentale impose un mode opératoire *in vitro* pour empêcher l'évolution du contexte expérimental vers la phase de résolution de l'inflammation.

Par ailleurs il etait important de ne pas seulement étudier l'action d'une seule cytokine. En effet nous savons combien l'action des cytokines sur l'hépatocyte est complexe suivant qu'elles agissent avec synergie ou antagonisme (cf. page 18). Aussi avons-nous préféré soumettre l'hépatocyte à un milieu conditionné, *i.e.* comprenant un mélange de cytokines.

Enfin parmi les deux lignées d'hépatomes humains disponibles, HepG2 et Hep3B, seules la dernière répondait aux critères imposés par nos conditions expérimentales [236]. Il a en effet été montré que HepG2 exprimait fortement et de façon constitutive le facteur de transcription C-EBP [237] qui à l'évidence induit un biais dans l'analyse de l'action de cytokines inflammatoires sur la cellule (cf. page 24). Ainsi l'ensemble de nos résultats montre une évolution normale des phénomènes connus liés au maintient du système en phase aiguë tel que par exemple l'absence de modification significative des proteines SOCS [85] (cf. page 32) qui sont des agents majeurs du retour vers l'état homéostasique.

6.4.2 Modifications négatives du schéma fonctionnel de l'hépatocyte

En réponse aux stimuli appliqués par les cytokines pro-inflammatoires du milieu conditionné, l'hépatocyte modifie son schéma fonctionnel. Les résultats de nos travaux montrent à l'évidence un contrôle négatif précoce mais éphémère de la transcription, révélé singulièrement par les altérations de l'abondance des ARNm des facteurs de transcription. La conséquence est une régulation de la phase traductionnelle au sein de l'hépatocyte avec notament la diminution significative de la traduction de protéines majoritairement impliquées dans les processus de détoxification et dans le métabolisme hépatocytaire. La modification de la traduction de ces protéines disparait en fin de phase aiguë malgré l'abcence de retour à l'homéostasie.
L'exemple est donné par la régulation négative de l'alcool-déshydrogénase et des enzymes importantes du métabolisme du glucose telle la glucose-6-phosphatase, du métabolisme des acides gras telle la stéanoyl-CoA désaturase et du métabolisme du cholestérol telle la 24-déshydrocholestérol désaturase. L'inhibition de ces voies métaboliques est en corrélation avec d'une part la répression du facteur de transcription HNF-4 qui est un activateur de métabolisme hépatocytaire [238] et l'activation du facteur de transcription STAT-3 qui est un répresseur de la glucogénèse dépendante à la glucose-6-phostase [239].

Nous éclairons alors d'un nouveau jour la séquence des événements qui constituent la phase aiguë de l'inflammation. Ainsi la période de la phase aiguë durant laquelle l'hépatocyte augmente la traduction des APP est postérieure au phénomène précédemment décrit de diminution de la traduction de protéines impliquées dans les processus de détoxification et dans le métabolisme hépatocytaire. Si cette diminution n'est pas dûe au retour à l'état homéostatique interdit par nos conditions expérimentales, il est fort probable qu'elle corresponde à une stratégie d'économie des dépenses en acides-aminés [20].

6.4.3 Stabilité des ARNm

Contrairement à d'autre travaux qui ont estimé le rôle de la stabilité des ARNm dans le jeu des régulations par déductions des analyses combinées du taux de transcription et de l'abondance des ARNm [240, 241], nous avons établi une méthode d'estimation directe de la valeur de la stabilité des ARNm.
À chaque temps d'adjonction d'actinomycine D et à chaque durée de stimulation en milieu conditionné, le rapport d'abondance en ARNm entre les cellules situées en milieu conditionné et celles situées en milieu non-conditionné a été mesuré. Pour chaque ARNm une estimation du coefficient β de régression linéaire a été calculé et représente la vitesse relative de dégradation de l'ARNm, *i.e.* la stabilité des ARNm.

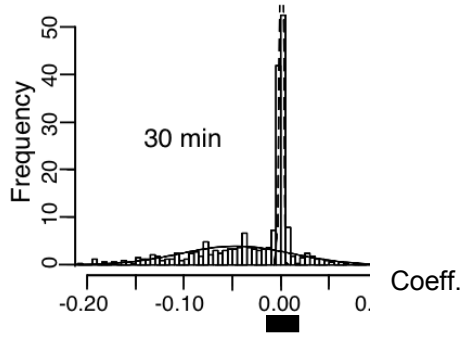


FIG. 6.2 – *Distributions des stabilités des ARNm.* L'abscisse représente les valeurs de coefficients de régressions linéaires (β) pour chacun des ARNm au cours des mesures d'abondance relevées à chaque adjonction d'actinomycine D. Deux distributions de fonctions de loi normale semblent être mélangées ($N_0(\mu_0 = 0, \sigma_0)$ et $N_1(\mu_1 = -0.05, \sigma_1)$). Après estimation des paramètres de ces fonctions, il est possible de déterminer un intervalle de confiance (rectangle noir) au delà duquel les coefficients sont significativement différents de zéro.

L'observation des distributions des β calculés pour chaque ARNm permet de suspecter le mélange de deux populations de coefficients, chacun distribué selon une loi normale (cf figure 6.2) :

- une première représentant les coefficients nuls ou assimilés nuls et attribués aux ARNm se dégradant à la même vitesse dans les cellules en milieu conditionné que dans celles en milieu non-conditionné ;
- une seconde représentant les coefficients non nuls et attribués aux ARNm ne se dégradant pas à la même vitesse dans les cellules en milieu conditionné que dans celles en milieu non-conditionné.

Afin d'estimer les fontières entre les β nuls et les non-nuls, il fallait determiner les paramètres des fonctions de distribution des deux lois normales $N(\mu, \sigma)$. Pour cela un programme réalisé en R s'exécute de façon itérative jusqu'à convergence du maximum de vraissemblance (cf. page 86). Les paramètres des lois normales ainsi estimés, un *t-test* permet d'estimer au risque α les seuils de confiance au delà desquels les coefficients β sont non-nuls. De cette manière, nous avons établi la liste des ARNm ne se dégradant pas à la même vitesse dans les cellules placée en milieu conditionné que dans celles placée en milieu non-conditionné.

Nos travaux montrent alors pour la première fois la prédominance d'un contrôle de la stabilité des ARNm comme moyen de régulation du métabolisme des hépatocytes au cours de la phase aiguë de l'inflammation et en particulier révèle que nombre de ces ARNm subissent une accélération de leur dégradation donc une diminution de leur stabilité. Ces résultats sont par ailleurs corroborés par deux autres observations issues de notre travail :

- les modifications positives et précoces de l'abondance des ribonucléoprotéines nucléaires dont notamment la hnRNP-D qui contrôle directement la stabilité des ARNm [242] ;
- les modifications précoces de l'abondance des enzymes intervenant au cours de la voie des MAPK dont le principal effet est une repression des MAPK et en conséquence une diminution de la stabilité des ARNm [243].

Il a de plus été démontré que certaines caractéristiques de la séquence de l'ARNm influent sur sa stabilité [88], notamment :

- la longeur de la séquence non-traduite 3'-UTR. Ainsi comme attendu dans notre étude l'observation des longueurs des séquences 3'-UTR des ARNm, révèle une longueur significativement plus courte pour les ARNm régulés que pour les ARNm non régulés ;
- l'abondance des motifs *AU-rich elements* (ARE). Les fréquences d'apparition des différents motifs ARE connus à ce jour [244] sont plus faibles dans les séquences d'ARNm régulés que dans celles de ceux n'étant pas régulés. Bien que cette caractéristique soit un facteur essentiel de la stabilité des

ARNm, elle ne permet cependant pas de prédire l'orientation positive ou négative de la régulation de la stabilité [243]

La perte précoce de la stabilité au cours de la phase aiguë de l'inflammation pour une majorité d'ARNm est en accord avec la répression éphémère et précoce de la transcription. Enfin la phase aiguë de l'inflammation est une période particulièrement active et il est raisonnable de penser que pour éviter un emballement de la réaction inflammatoire la cellule puisse limiter l'abondance de certains ARNm dont la présence persistente pourrait être délétère.

Comme attendu nombre d'APP et d'APRIP subissent une régulation transcriptionnelle durant la phase aiguë de l'inflammation notament une activation dans la période précoce et une répression dans la période tardive. Afin d'observer ce phénomène, nous avons été les premiers à étendre la technique de Run-On vers une dimension transcriptomique et avons relevé d'une part parmi un grand nombre de facteurs de transcription une corrélation entre l'activation de la transcription et l'abondance de leur ARNm propre mais d'autre part une activation de certains facteurs de transcription qui peut intervenir différemment sur la régulation des autres protéines suivant un effet activateur ou régulateur.

Ce n'est pas la première fois que des travaux montrent qu'une activation de la transcription ou une augmentation de la stabilité peut être opposée à l'abondance d'ARNm ou même au niveau de traduction [88, 240, 241]. Notre étude révèle ainsi dans ses dimensions transcriptomique et cinétique la séquence d'évènements de la phase aiguë de la transcription jusqu'à la traduction et montre par ailleurs combien l'hépatocyte placé en contexte de phase aiguë de l'inflammation peut réagir en synergie ou en antagonisme sur ses leviers de régulations que sont la transcription et la stabilité afin d'obtenir l'abondance d'ARNm nécessaire pour sa réponse au stimulus cytokinique.

Chapitre 7

Conclusion

Dès 1999 notre équipe a souhaité se doter d'une plateforme de puces à ADN spécialisée pour l'étude de l'expression génique hépatique chez l'humain au cours de la phase aiguë de l'inflammation systémique mais également au cours de carcinomes hépatocellulaires [245] ou du développement [246].

Une première partie de mon travail présenté dans la revue *Hepatology* en 2004 [236] fût donc pour l'essentiel d'une part la préparation et la validation d'une plate-forme « puce à ADN » et d'autre part la recherche de marqueurs de l'inflammation aiguë systémique potentiellement plasmatiques.
L'outil ainsi développé repose sur trois éléments fondamentaux :

- une puce à ADN nommée Liverpool et dédiée au foie humain. Le choix des sondes déposées sur la puce et leur nombre nous autorise un regard sur l'ensemble du transcriptome hépatique humain ;
- une logistique informatique. Constituée d'une part d'un serveur de base de données actualisées et sauvegardées périodiquement et automatiquement, et d'autre part d'outils d'analyses statistiques des données provenant des expériences, LiverTools est un ensemble d'outils placés dans un réseau informatique au coeur du laboratoire dont le but est d'appréhender et de faciliter le travail du chercheur qui utilise Liverpool ;
- un savoir-faire de haute qualité tant au niveau de la préparation des expériences que de leurs analyses et détenu par l'ensemble du personnel de l'équipe.

Ainsi l'utilisation de cet outil dans le cadre d'une étude comparative entre des prélévements de tissus hépatiques inflammatoires et des prélèvements de tissus hépatiques sains, et associée à une étude des facteurs cliniques des patients, nous a permi de mettre en lumière les principaux acteurs du processus inflammatoire chez l'humain. Le travail met particulièrement en avant la découverte de 134 marqueurs du niveau d'intensité de l'état inflammatoire révélant alors d'une part que 5 d'entre eux sont très probablement des protéines plasmatiques et d'autre part qu'il existe une forte corrélation entre le niveau d'expression des APP positives et le niveau d'intensité de l'état inflammatoire.

Le seconde partie de mon travail fût la détermination de la mise en scène de l'ensemble des acteurs de la phase aiguë. Le travail publié dans la revue *Hepatology* en 2005 [247], nous permet d'identifier plusieurs cascades d'événements dans la lignée de cellules d'hépatome Hep3B placées en culture et stimulées par des cytokines pro-inflammatoires gardées ainsi de ce fait en phase aiguë artificiellement prolongée. En particulier, la répression transitoire d'abondance des ARNm codant les protéines du métabolisme hépatique est un événement marquant. Notre étude a aussi, pour la première fois, évalué l'impact de régulations modifiant transitoirement la transcription des gènes, la stabilité des ARNm, ou leur niveau de

traduction, et a montré que la perte de stabilité des ARNm est un événement prédominant lors de la réponse de l'hépatocyte à l'inflammation aiguë systémique.

Ce travail montre combien les puces à ADN associées à d'autres techniques telles la Q-RT-PCR et le RunON, sont un outil particulièrement bien adapté à la compréhension fine d'un évènement biologique en mettant en lumière l'ensemble des interactions entres les nombreux gènes protagonistes.
L'interprétation des résultats reste cependant autant un art qu'une science et ce, malgré l'émergence ces dernières années de standards indispensables qui permettent d'orienter l'expérimentateur vers une interprétation des résultats plus objective et fiable.
En effet des protocoles expérimentaux plus reproductibles, des annotations plus fiables, des méthodes standardisées d'enregistrement des données, la définition d'ontologies et des modèles mathématiques particulièrement bien adaptés nous permettront d'accroître davantage nos capacités d'interprétation des données issues des puces à ADN.

Table des figures

Liste des tableaux

Bibliographie

[1] J. J. Gumucio, « Functional and anatomic heterogeneity in the liver acinus : impact on transport », *Am J Physiol*, vol. 244, no. 6, p. G578–82, 1983.

[2] Z. Kmiec, « Cooperation of liver cells in health and disease », *Adv Anat Embryol Cell Biol*, vol. 161, p. 1–151, 2001.

[3] I. N. Crispe, « Hepatic t cells and liver tolerance », *Nat Rev Immunol*, vol. 3, no. 1, p. 51–62, 2003.

[4] H. Senoo, « Structure and function of hepatic stellate cells », *Med Electron Microsc*, vol. 37, no. 1, p. 3–15, 2004.

[5] M. L. Hautekeete et A. Geerts, « The hepatic stellate (ito) cell : its role in human liver disease », *Virchows Arch*, vol. 430, no. 3, p. 195–207, 1997.

[6] M. Emoto et S. H. Kaufmann, « Liver nkt cells : an account of heterogeneity », *Trends Immunol*, vol. 24, no. 7, p. 364–9, 2003.

[7] M. Bollen, S. Keppens et W. Stalmans, « Specific features of glycogen metabolism in the liver », *Biochem J*, vol. 336 (Pt 1), p. 19–31, 1998.

[8] C. H. Lee, P. Olson et R. M. Evans, « Minireview : lipid metabolism, metabolic diseases, and peroxisome proliferator-activated receptors », *Endocrinology*, vol. 144, no. 6, p. 2201–7, 2003.

[9] D. M. Grant, « Detoxification pathways in the liver », *J Inherit Metab Dis*, vol. 14, no. 4, p. 421–30, 1991.

[10] C. O'Farrelly et I. N. Crispe, « Prometheus through the looking glass : reflections on the hepatic immune system », *Immunol Today*, vol. 20, no. 9, p. 394–8, 1999.

[11] P. A. Knolle et G. Gerken, « Local control of the immune response in the liver », *Immunol Rev*, vol. 174, p. 21–34, 2000.

[12] H. Baumann et J. Gauldie, « The acute phase response », *Immunol Today*, vol. 15, no. 2, p. 74–80, 1994.

[13] E. Olivier, E. Soury, J. L. Risler, F. Smih, K. Schneider, K. Lochner, J. Y. Jouzeau, G. H. Fey et J. P. Salier, « A novel set of hepatic mrnas preferentially expressed during an acute inflammation in rat represents mostly intracellular proteins », *Genomics*, vol. 57, no. 3, p. 352–64, 1999.

[14] G. H. Fey et G. M. Fuller, « Regulation of acute phase gene expression by inflammatory mediators », *Mol Biol Med*, vol. 4, no. 6, p. 323–38, 1987.

[15] I. Kushner, « Regulation of the acute phase response by cytokines », *Perspect Biol Med*, vol. 36, no. 4, p. 611–22, 1993.

[16] R. C. Bone, « Immunologic dissonance : a continuing evolution in our understanding of the systemic inflammatory response syndrome (sirs) and the multiple organ dysfunction syndrome (mods) », *Ann Intern Med*, vol. 125, no. 8, p. 680–7, 1996.

[17] E. Metchnikoff, *Immunity in infectious deseases (1905).* New York : Johnson reprint corporation, 1968.

[18] R. Fauve et M.-B. Hevin, « Réaction inflammatoire et réactions immunitaires », *in L'inflammation* (J. L. Eurotext, éd.), p. 10–20, 1998.

[19] J. A. Hoffmann, J. M. Reichhart et C. Hetru, « Innate immunity in higher insects », *Curr Opin Immunol*, vol. 8, no. 1, p. 8–13, 1996.

[20] C. Gabay et I. Kushner, « Acute-phase proteins and other systemic responses to inflammation », *N Engl J Med*, vol. 340, no. 6, p. 448–54, 1999.

[21] E. Lolis et R. Bucala, « Therapeutic approaches to innate immunity : severe sepsis and septic shock », *Nat Rev Drug Discov*, vol. 2, no. 8, p. 635–45, 2003.

[22] W. Kiefer et G. Dannhardt, « Cox-2 inhibition and the control of pain », *Curr Opin Investig Drugs*, vol. 3, no. 9, p. 1348–58, 2002.

[23] A. Koj, « Initiation of acute phase response and synthesis of cytokines », *Biochim Biophys Acta*, vol. 1317, no. 2, p. 84–94, 1996.

[24] C. A. Dinarello, « Proinflammatory cytokines », *Chest*, vol. 118, no. 2, p. 503–8, 2000.

[25] G. Ramadori et B. Christ, « Cytokines and the hepatic acute-phase response », *Semin Liver Dis*, vol. 19, no. 2, p. 141–55, 1999.

[26] H. Moshage, « Cytokines and the hepatic acute phase response », *J Pathol*, vol. 181, no. 3, p. 257–66, 1997.

[27] S. M. Opal et V. A. DePalo, « Anti-inflammatory cytokines », *Chest*, vol. 117, no. 4, p. 1162–72, 2000.

[28] G. Ramadori et T. Armbrust, « Cytokines in the liver », *Eur J Gastroenterol Hepatol*, vol. 13, no. 7, p. 777–84, 2001.

[29] T. Lawrence, D. A. Willoughby et D. W. Gilroy, « Anti-inflammatory lipid mediators and insights into the resolution of inflammation », *Nat Rev Immunol*, vol. 2, no. 10, p. 787–95, 2002.

[30] A. Mackiewicz, T. Speroff, M. K. Ganapathi et I. Kushner, « Effects of cytokine combinations on acute phase protein production in two human hepatoma cell lines », *J Immunol*, vol. 146, no. 9, p. 3032–7, 1991.

[31] B. E. Barton, « Il-6 : insights into novel biological activities », *Clin Immunol Immunopathol*, vol. 85, no. 1, p. 16–20, 1997.

[32] J. M. Cavaillon, « Pro- versus anti-inflammatory cytokines : myth or reality », *Cell Mol Biol (Noisy-le-grand)*, vol. 47, no. 4, p. 695–702, 2001.

[33] C. A. Dinarello, « Biologic basis for interleukin-1 in disease », *Blood*, vol. 87, no. 6, p. 2095–147, 1996.

[34] R. Volpes, J. J. van den Oord, R. De Vos et V. J. Desmet, « Hepatic expression of type a and type b receptors for tumor necrosis factor », *J Hepatol*, vol. 14, no. 2-3, p. 361–9, 1992.

[35] K. Pfeffer, T. Matsuyama, T. M. Kundig, A. Wakeham, K. Kishihara, A. Shahinian, K. Wiegmann, P. S. Ohashi, M. Kronke et T. W. Mak, « Mice deficient for the 55 kd tumor necrosis factor receptor are resistant to endotoxic shock, yet succumb to l. monocytogenes infection », *Cell*, vol. 73, no. 3, p. 457–67, 1993.

[36] T. Kishimoto, S. Akira, M. Narazaki et T. Taga, « Interleukin-6 family of cytokines and gp130 », *Blood*, vol. 86, no. 4, p. 1243–54, 1995.

[37] J. Darnell, J. E., I. M. Kerr et G. R. Stark, « Jak-stat pathways and transcriptional activation in response to ifns and other extracellular signaling proteins », *Science*, vol. 264, no. 5164, p. 1415–21, 1994.

[38] P. C. Heinrich, I. Behrmann, S. Haan, H. M. Hermanns, G. Muller-Newen et F. Schaper, « Principles of interleukin (il)-6-type cytokine signalling and its regulation », *Biochem J*, vol. 374, no. Pt 1, p. 1–20, 2003.

[39] I. M. Verma, J. K. Stevenson, E. M. Schwarz, D. Van Antwerp et S. Miyamoto, « Rel/nf-kappa b/i kappa b family : intimate tales of association and dissociation », *Genes Dev*, vol. 9, no. 22, p. 2723–35, 1995.

[40] Q. Li et I. M. Verma, « Nf-kappab regulation in the immune system », *Nat Rev Immunol*, vol. 2, no. 10, p. 725–34, 2002.

[41] K. Burns, F. Martinon, C. Esslinger, H. Pahl, P. Schneider, J. L. Bodmer, F. Di Marco, L. French et J. Tschopp, « Myd88, an adapter protein involved in interleukin-1 signaling », *J Biol Chem*, vol. 273, no. 20, p. 12203–9, 1998.

[42] L. E. Jensen, M. Muzio, A. Mantovani et A. S. Whitehead, « Il-1 signaling cascade in liver cells and the involvement of a soluble form of the il-1 receptor accessory protein », *J Immunol*, vol. 164, no. 10, p. 5277–86, 2000.

[43] Q. Huang, J. Yang, Y. Lin, C. Walker, J. Cheng, Z. G. Liu et B. Su, « Differential regulation of interleukin 1 receptor and toll-like receptor signaling by mekk3 », *Nat Immunol*, vol. 5, no. 1, p. 98–103, 2004.

[44] G. Chen et D. V. Goeddel, « Tnf-r1 signaling : a beautiful pathway », *Science*, vol. 296, no. 5573, p. 1634–5, 2002.

[45] M. Karin et Y. Ben-Neriah, « Phosphorylation meets ubiquitination : the control of nf-[kappa]b activity », *Annu Rev Immunol*, vol. 18, p. 621–63, 2000.

[46] D. M. Rothwarf, E. Zandi, G. Natoli et M. Karin, « Ikk-gamma is an essential regulatory subunit of the ikappab kinase complex », *Nature*, vol. 395, no. 6699, p. 297–300, 1998.

[47] K. Shuai et B. Liu, « Regulation of jak-stat signalling in the immune system », *Nat Rev Immunol*, vol. 3, no. 11, p. 900–11, 2003.

[48] J. Darnell, J. E., « Stats and gene regulation », *Science*, vol. 277, no. 5332, p. 1630–5, 1997.

[49] D. Anhuf, M. Weissenbach, J. Schmitz, R. Sobota, H. M. Hermanns, S. Radtke, S. Linnemann, I. Behrmann, P. C. Heinrich et F. Schaper, « Signal transduction of il-6, leukemia-inhibitory factor, and oncostatin m : structural receptor requirements for signal attenuation », *J Immunol*, vol. 165, no. 5, p. 2535–43, 2000.

[50] K. Shuai, C. M. Horvath, L. H. Huang, S. A. Qureshi, D. Cowburn et J. Darnell, J. E., « Interferon activation of the transcription factor stat91 involves dimerization through sh2-phosphotyrosyl peptide interactions », *Cell*, vol. 76, no. 5, p. 821–8, 1994.

[51] S. Yamada, S. Shiono, A. Joo et A. Yoshimura, « Control mechanism of jak/stat signal transduction pathway », *FEBS Lett*, vol. 534, no. 1-3, p. 190–6, 2003.

[52] J. M. Kyriakis et J. Avruch, « Mammalian mitogen-activated protein kinase signal transduction pathways activated by stress and inflammation », *Physiol Rev*, vol. 81, no. 2, p. 807–69, 2001.

[53] E. Herlaar et Z. Brown, « p38 mapk signalling cascades in inflammatory disease », *Mol Med Today*, vol. 5, no. 10, p. 439–47, 1999.

[54] G. Zhou, Z. Q. Bao et J. E. Dixon, « Components of a new human protein kinase signal transduction pathway », *J Biol Chem*, vol. 270, no. 21, p. 12665–9, 1995.

[55] C. A. Hazzalin et L. C. Mahadevan, « Mapk-regulated transcription : a continuously variable gene switch ? », *Nat Rev Mol Cell Biol*, vol. 3, no. 1, p. 30–40, 2002.

[56] V. Ossipow, P. Descombes et U. Schibler, « Ccaat/enhancer-binding protein mrna is translated into multiple proteins with different transcription activation potentials », *Proc Natl Acad Sci U S A*, vol. 90, no. 17, p. 8219–23, 1993.

[57] J. Lekstrom-Himes et K. G. Xanthopoulos, « Biological role of the ccaat/enhancer-binding protein family of transcription factors », *J Biol Chem*, vol. 273, no. 44, p. 28545–8, 1998.

[58] D. P. Ramji et P. Foka, « Ccaat/enhancer-binding proteins : structure, function and regulation », *Biochem J*, vol. 365, no. Pt 3, p. 561–75, 2002.

[59] V. Poli, « The role of c/ebp isoforms in the control of inflammatory and native immunity functions », *J Biol Chem*, vol. 273, no. 45, p. 29279–82, 1998.

[60] R. Kolesnick et D. W. Golde, « The sphingomyelin pathway in tumor necrosis factor and interleukin-1 signaling », *Cell*, vol. 77, no. 3, p. 325–8, 1994.

[61] M. R. Ruocco, X. Chen, C. Ambrosino, E. Dragonetti, W. Liu, M. Mallardo, G. De Falco, C. Palmieri, G. Franzoso, I. Quinto, S. Venuta et G. Scala, « Regulation of hiv-1 long terminal repeats by interaction of c/ebp(nf-il6) and nf-kappab/rel transcription factors », *J Biol Chem*, vol. 271, no. 37, p. 22479–86, 1996.

[62] K. P. LeClair, M. A. Blanar et P. A. Sharp, « The p50 subunit of nf-kappa b associates with the nf-il6 transcription factor », *Proc Natl Acad Sci U S A*, vol. 89, no. 17, p. 8145–9, 1992.

[63] C. Xia, J. K. Cheshire, H. Patel et P. Woo, « Cross-talk between transcription factors nf-kappa b and c/ebp in the transcriptional regulation of genes », *Int J Biochem Cell Biol*, vol. 29, no. 12, p. 1525–39, 1997.

[64] C. A. Cantwell, E. Sterneck et P. F. Johnson, « Interleukin-6-specific activation of the c/ebpdelta gene in hepatocytes is mediated by stat3 and sp1 », *Mol Cell Biol*, vol. 18, no. 4, p. 2108–17, 1998.

[65] I. Kushner et A. Mackiewicz, « Acute phase proteins as disease markers », *Dis Markers*, vol. 5, no. 1, p. 1–11, 1987.

[66] M. Scotte, M. Hiron, S. Masson, S. Lyoumi, F. Banine, P. Teniere, J. P. Lebreton et M. Daveau, « Differential expression of cytokine genes in monocytes, peritoneal macrophages and liver following endotoxin- or turpentine-induced inflammation in rat », *Cytokine*, vol. 8, no. 2, p. 115–20, 1996.

[67] E. Soury, E. Olivier, D. Simon, P. Ruminy, K. Kitada, M. Hiron, M. Daveau, Y. Boyd, T. Serikawa, J. L. Guenet et J. P. Salier, « Chromosomal assignments of mammalian genes with an acute inflammation-regulated expression in liver », *Immunogenetics*, vol. 53, no. 8, p. 634–42, 2001.

[68] T. W. Du Clos, « Function of c-reactive protein », *Ann Med*, vol. 32, no. 4, p. 274–8, 2000.

[69] S. Urieli-Shoval, R. P. Linke et Y. Matzner, « Expression and function of serum amyloid a, a major acute-phase protein, in normal and disease states », *Curr Opin Hematol*, vol. 7, no. 1, p. 64–9, 2000.

[70] S. K. Lim, B. Ferraro, K. Moore et B. Halliwell, « Role of haptoglobin in free hemoglobin metabolism », *Redox Rep*, vol. 6, no. 4, p. 219–27, 2001.

[71] S. Janciauskiene, « Conformational properties of serine proteinase inhibitors (serpins) confer multiple pathophysiological roles », *Biochim Biophys Acta*, vol. 1535, no. 3, p. 221–35, 2001.

[72] R. Engler, « [acute-phase proteins in inflammation] », *C R Seances Soc Biol Fil*, vol. 189, no. 4, p. 563–78, 1995.

[73] M. Ombrellino, H. Wang, H. Yang, M. Zhang, J. Vishnubhakat, A. Frazier, L. A. Scher, S. G. Friedman et K. J. Tracey, « Fetuin, a negative acute phase protein, attenuates tnf synthesis and the innate inflammatory response to carrageenan », *Shock*, vol. 15, no. 3, p. 181–5, 2001.

[74] H. Wang, M. Zhang, M. Bianchi, B. Sherry, A. Sama et K. J. Tracey, « Fetuin (alpha2-hs-glycoprotein) opsonizes cationic macrophagedeactivating molecules », *Proc Natl Acad Sci U S A*, vol. 95, no. 24, p. 14429–34, 1998.

[75] W. Y. Almawi, H. N. Beyhum, A. A. Rahme et M. J. Rieder, « Regulation of cytokine and cytokine receptor expression by glucocorticoids », *J Leukoc Biol*, vol. 60, no. 5, p. 563–72, 1996.

[76] C. N. Serhan, « Lipoxin biosynthesis and its impact in inflammatory and vascular events », *Biochim Biophys Acta*, vol. 1212, no. 1, p. 1–25, 1994.

[77] D. W. Gilroy, P. R. Colville-Nash, D. Willis, J. Chivers, M. J. Paul-Clark et D. A. Willoughby, « Inducible cyclooxygenase may have anti-inflammatory properties », *Nat Med*, vol. 5, no. 6, p. 698–701, 1999.

[78] C. A. Dinarello, « Interleukin-1, interleukin-1 receptors and interleukin-1 receptor antagonist », *Int Rev Immunol*, vol. 16, no. 5-6, p. 457–99, 1998.

[79] S. Bevan et J. G. Raynes, « Il-1 receptor antagonist regulation of acute phase protein synthesis in human hepatoma cells », *J Immunol*, vol. 147, no. 8, p. 2574–8, 1991.

[80] C. Gabay, B. Genin, G. Mentha, P. B. Iynedjian, P. Roux-Lombard et P. A. Guerne, « Il-1 receptor antagonist (il-1ra) does not inhibit the production of c-reactive protein or serum amyloid a protein by human primary hepatocytes. differential regulation in normal and tumour cells », *Clin Exp Immunol*, vol. 100, no. 2, p. 306–13, 1995.

[81] A. Koj, « Termination of acute-phase response : role of some cytokines and anti-inflammatory drugs », *Gen Pharmacol*, vol. 31, no. 1, p. 9–18, 1998.

[82] Z. Xing, J. Gauldie, G. Cox, H. Baumann, M. Jordana, X. F. Lei et M. K. Achong, « Il-6 is an antiinflammatory cytokine required for controlling local or systemic acute inflammatory responses », *J Clin Invest*, vol. 101, no. 2, p. 311–20, 1998.

[83] S. M. Opal, J. C. Wherry et P. Grint, « Interleukin-10 : potential benefits and possible risks in clinical infectious diseases », *Clin Infect Dis*, vol. 27, no. 6, p. 1497–507, 1998.

[84] S. Wormald et D. J. Hilton, « Inhibitors of cytokine signal transduction », *J Biol Chem*, vol. 279, no. 2, p. 821–4, 2004.

[85] X. Wu, D. N. Herndon et S. E. Wolf, « Growth hormone down-regulation of interleukin-1beta and interleukin-6 induced acute phase protein gene expression is associated with increased gene expression of suppressor of cytokine signal-3 », *Shock*, vol. 19, no. 4, p. 314–20, 2003.

[86] P. Ruminy, C. Gangneux, S. Claeyssens, M. Scotte, M. Daveau et J. P. Salier, « Gene transcription in hepatocytes during the acute phase of a systemic inflammation : from transcription factors to target genes », *Inflamm Res*, vol. 50, no. 8, p. 383–90, 2001.

[87] S. L. Jiang, D. Samols, D. Rzewnicki, S. S. Macintyre, I. Greber, J. Sipe et I. Kushner, « Kinetic modeling and mathematical analysis indicate that acute phase gene expression in hep 3b cells is regulated by both transcriptional and posttranscriptional mechanisms », *J Clin Invest*, vol. 95, no. 3, p. 1253–61, 1995.

[88] E. Yang, E. van Nimwegen, M. Zavolan, N. Rajewsky, M. Schroeder, M. Magnasco et J. Darnell, J. E., « Decay rates of human mrnas : correlation with functional characteristics and sequence attributes », *Genome Res*, vol. 13, no. 8, p. 1863–72, 2003.

[89] C. Y. Chen et A. B. Shyu, « Au-rich elements : characterization and importance in mrna degradation », *Trends Biochem Sci*, vol. 20, no. 11, p. 465–70, 1995.

[90] J. D. Keene, « Ribonucleoprotein infrastructure regulating the flow of genetic information between the genome and the proteome », *Proc Natl Acad Sci U S A*, vol. 98, no. 13, p. 7018–24, 2001.

[91] F. Maurer et R. L. Medcalf, « Plasminogen activator inhibitor type 2 gene induction by tumor necrosis factor and phorbol ester involves transcriptional and post-transcriptional events. identification of a functional nonameric au-rich motif in the 3'-untranslated region », *J Biol Chem*, vol. 271, no. 42, p. 26074–80, 1996.

[92] M. A. Frevel, T. Bakheet, A. M. Silva, J. G. Hissong, K. S. Khabar et B. R. Williams, « p38 mitogen-activated protein kinase-dependent and -independent signaling of mrna stability of au-rich element-containing transcripts », *Mol Cell Biol*, vol. 23, no. 2, p. 425–36, 2003.

[93] M. Schena, D. Shalon, R. W. Davis et P. O. Brown, « Quantitative monitoring of gene expression patterns with a complementary dna microarray », *Science*, vol. 270, no. 5235, p. 467–70, 1995.

[94] H. Dong, N. Toyoda, H. Yoneyama, M. Kurachi, T. Kasahara, Y. Kobayashi, H. Inadera, S. Hashimoto et K. Matsushima, « Gene expression profile analysis of the mouse liver during bacteria-induced fulminant hepatitis by a cdna microarray system », *Biochem Biophys Res Commun*, vol. 298, no. 5, p. 675–86, 2002.

[95] N. Yano, N. A. Habib, K. J. Fadden, H. Yamashita, R. Mitry, H. Jauregui, A. Kane, M. Endoh et A. Rifai, « Profiling the adult human liver transcriptome : analysis by cdna array hybridization », *J Hepatol*, vol. 35, no. 2, p. 178–86, 2001.

[96] A. M. Chinnaiyan, M. Huber-Lang, C. Kumar-Sinha, T. R. Barrette, S. Shankar-Sinha, V. J. Sarma, V. A. Padgaonkar et P. A. Ward, « Molecular signatures of sepsis : multiorgan gene expression profiles of systemic inflammation », *Am J Pathol*, vol. 159, no. 4, p. 1199–209, 2001.

[97] J. P. Cobb, J. M. Laramie, G. D. Stormo, J. J. Morrissey, W. D. Shannon, Y. Qiu, I. E. Karl, T. G. Buchman et R. S. Hotchkiss, « Sepsis gene expression profiling : murine splenic compared with hepatic responses determined by using complementary dna microarrays », *Crit Care Med*, vol. 30, no. 12, p. 2711–21, 2002.

[98] M. A. Higgins, B. R. Berridge, B. J. Mills, A. E. Schultze, H. Gao, G. H. Searfoss, T. K. Baker et T. P. Ryan, « Gene expression analysis of the acute phase response using a canine microarray », *Toxicol Sci*, vol. 74, no. 2, p. 470–84, 2003.

[99] C. Fang, S. Yoon, N. Tindberg, H. A. Jarvelainen, K. O. Lindros et M. Ingelman-Sundberg, « Hepatic expression of multiple acute phase proteins and down-regulation of nuclear receptors after acute endotoxin exposure », *Biochem Pharmacol*, vol. 67, no. 7, p. 1389–97, 2004.

[100] M. R. Dasu, J. P. Cobb, J. M. Laramie, T. P. Chung, M. Spies et R. E. Barrow, « Gene expression profiles of livers from thermally injured rats », *Gene*, vol. 327, no. 1, p. 51–60, 2004.

[101] ANSI/X3/SPARC, « American national standard institute study group on dbms : Interim report », rap. tech., Bulletin of the ACM SIGMOD, 1975.

[102] A. Cardon et C. Dabancourt, *Initiation à l'algorithmique objet*. eroylles éd., 2001.

[103] S. Alhir, *Introduction à UML*. o'reilly éd., 2004.

[104] L. Wang, P. Rodriguez-Tome, N. Redaschi, P. McNeil, A. Robinson et P. Lijnzaad, « Accessing and distributing embl data using corba (common object request broker architecture) », *Genome Biol*, vol. 1, 2000.

[105] R. Dowell, R. Jokerst, A. Day, S. Eddy et L. Stein, « The distributed annotation system. », *BMC Bioinformatics*, vol. 2, p. 7, 2001.

[106] H. Tardieu, A. Rochfeld et R. Colleti, *La méthode MERISE tome1*. les éditions d'organisation éd., 1983.

[107] G. Booch, J. Rumbaugh et I. Jacobson, *The Unified Modeling Language User Guide*. addison-wesley éd., 1997.

[108] C. F. Taylor, N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, E. W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M. J. Deery, J. A. Howard, T. Dunkley, R. Aebersold, D. B. Kell, K. S. Lilley, P. Roepstorff, r. Yates, J. R., A. Brass, A. J. Brown, P. Cash, S. J. Gaskell, S. J. Hubbard et S. G. Oliver, « A systematic approach to modeling, capturing, and disseminating proteomics experimental data », *Nat Biotechnol*, vol. 21, no. 3, p. 247–54, 2003.

[109] C. Soutou, *De SQL à UML*. eyrolles éd., 2002.

[110] S. Lawrence et C. L. Giles, « Accessibility of information on the web », *Nature*, vol. 400, no. 6740, p. 107–9, 1999.

[111] F. Achard, G. Vaysseix et E. Barillot, « Xml, bioinformatics and data integration », *Bioinformatics*, vol. 17, no. 2, p. 115–25, 2001.

[112] E. Harold et W. Means, *XML in a nutshell*. Paris : Oreilly, oreilly éd., 2002.

[113] S. M. Searle, J. Gilbert, V. Iyer et M. Clamp, « The otter annotation system », *Genome Res*, vol. 14, no. 5, p. 963–70, 2004.

[114] S. Xirasagar, S. Gustafson, B. A. Merrick, K. B. Tomer, S. Stasiewicz, D. D. Chan, r. Yost, K. J., r. Yates, J. R., S. Sumner, N. Xiao et M. D. Waters, « Cebs object model for systems biology data, sysbio-om », *Bioinformatics*, vol. 20, no. 13, p. 1–12, 2004.

[115] G. Gardarin, *Internet/intranet et bases de données*. eroylles éd., 1999.

[116] E.-F. Codd, « Derivability, redundancy and consistency of relations stored in large databanks », Rap. tech. RJ599, IBM, 19 août 1969 1969.

[117] P. Dubois, *MySQL*. New riders publishing, 2000.

[118] P. D. Karp, « Database links are a foundation for interoperability », *Trends Biotechnol*, vol. 14, no. 8, p. 273–9, 1996.

[119] M. F. Miles, « Microarrays : lost in a storm of data ? », *Nat Rev Neurosci*, vol. 2, no. 6, p. 441–3, 2001.

[120] r. Petricoin, E. F., J. L. Hackett, L. J. Lesko, R. K. Puri, S. I. Gutman, K. Chumakov, J. Woodcock, J. Feigal, D. W., K. C. Zoon et F. D. Sistare, « Medical applications of microarray technologies : a regulatory science perspective », *Nat Genet*, vol. 32 Suppl, p. 474–9, 2002.

[121] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield et E. S. Lander, « Molecular classification of cancer : class discovery and class prediction by gene expression monitoring », *Science*, vol. 286, no. 5439, p. 531–7, 1999.

[122] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, J., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown et L. M. Staudt, « Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling », *Nature*, vol. 403, no. 6769, p. 503–11, 2000.

[123] S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin et A. M. Chinnaiyan, « Delineation of prognostic biomarkers in prostate cancer », *Nature*, vol. 412, no. 6849, p. 822–6, 2001.

[124] T. O. Nielsen, R. B. West, S. C. Linn, O. Alter, M. A. Knowling, J. X. O'Connell, S. Zhu, M. Fero, G. Sherlock, J. R. Pollack, P. O. Brown, D. Botstein et M. van de Rijn, « Molecular characterisation of soft tissue tumours : a gene expression study », *Lancet*, vol. 359, no. 9314, p. 1301–7, 2002.

[125] M. H. Jones, C. Virtanen, D. Honjoh, T. Miyoshi, Y. Satoh, S. Okumura, K. Nakagawa, H. Nomura et Y. Ishikawa, « Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles », *Lancet*, vol. 363, no. 9411, p. 775–81, 2004.

[126] L. M. Staudt et P. O. Brown, « Genomic views of the immune system* », *Annu Rev Immunol*, vol. 18, p. 829–59, 2000.

[127] A. Brazma, A. Robinson, G. Cameron et M. Ashburner, « One-stop shop for microarray data », *Nature*, vol. 403, no. 6771, p. 699–700, 2000.

[128] J. Pontius, L. Wagner et S. G.D., « Unigene : a unified view of the transcriptome », *in The NCBI Handbook* (N. C. f. B. Information, éd.), p. 1–12, 2003.

[129] K. D. Pruitt, K. S. Katz, H. Sicotte et D. R. Maglott, « Introducing refseq and locuslink : curated human genome resources at the ncbi », *Trends Genet*, vol. 16, no. 1, p. 44–7, 2000.

[130] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout et M. Schneider, « The swiss-prot protein knowledgebase and its supplement trembl in 2003 », *Nucleic Acids Res*, vol. 3, no. 1, p. 365–70, 2003.

[131] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno et M. Hattori, « The kegg resource for deciphering the genome », *Nucleic Acids Res*, vol. 32, p. D277–80, 2004.

[132] MGED, « Reporting structure for biological investigations working groups », 2004. http ://www.mged.org/Workgroups/rsbi/rsbi.html.

[133] J. Quackenbush, « Data standards for 'omic' science », *Nat Biotechnol*, vol. 22, no. 5, p. 613–4, 2004.

[134] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo et M. Vingron, « Minimum information about a microarray experiment (miame)-toward standards for microarray data », *Nat Genet*, vol. 29, no. 4, p. 365–71, 2001.

[135] C. A. Ball, G. Sherlock, H. Parkinson, P. Rocca-Sera, C. Brooksbank, H. C. Causton, D. Cavalieri, T. Gaasterland, P. Hingamp, F. Holstege, M. Ringwald, P. Spellman, J. Stoeckert, C. J., J. E.

Stewart, R. Taylor, A. Brazma et J. Quackenbush, « Standards for microarray data », *Science*, vol. 298, no. 5593, p. 539, 2002.

[136] C. A. Ball, G. Sherlock, H. Parkinson, P. Rocca-Sera, C. Brooksbank, H. C. Causton, D. Cavalieri, T. Gaasterland, P. Hingamp, F. Holstege, M. Ringwald, P. Spellman, J. Stoeckert, C. J., J. E. Stewart, R. Taylor, A. Brazma et J. Quackenbush, « The underlying principles of scientific publication », *Bioinformatics*, vol. 18, no. 11, p. 1409, 2002.

[137] C. A. Ball, G. Sherlock, H. Parkinson, P. Rocca-Sera, C. Brooksbank, H. C. Causton, D. Cavalieri, T. Gaasterland, P. Hingamp, F. Holstege, M. Ringwald, P. Spellman, J. Stoeckert, C. J., J. E. Stewart, R. Taylor, A. Brazma et J. Quackenbush, « A guide to microarray experiments–an open letter to the scientific journals », *Lancet*, vol. 360, no. 9338, p. 1019 ; author reply 1019, 2002.

[138] « Coming to terms with microarrays », *Nature Genetics*, vol. 32, p. 333–334, 2002.

[139] « Microarray standards at last », *Nature*, vol. 419, p. 323, 2002.

[140] C. Ball, A. Brazma, H. Causton, S. Chervitz, R. Edgar, P. Hingamp, J. C. Matese, C. Icahn, H. Parkinson, J. Quackenbush, M. Ringwald, S. A. Sansone, G. Sherlock, P. Spellman, C. Stoeckert, Y. Tateno, R. Taylor, J. White et N. Winegarden, « An open letter on microarray data from the mged society », *Microbiology*, vol. 150, no. Pt 11, p. 3522–4, 2004.

[141] MGED, « An open letter to the scientific journals », 2004. http ://www.mged.org/MIAME_open_letter.html.

[142] R. Inpharmatics, « Gene expression markup language (geml) », 2000. http ://www.rosettabio.com/tech/geml/default.htm.

[143] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, J. Stoeckert, C. J. et A. Brazma, « Design and implementation of microarray gene expression markup language (mage-ml) », *Genome Biol*, vol. 3, no. 9, p. RESEARCH0046, 2002.

[144] M. Y. Galperin, « The molecular biology database collection : 2004 update », *Nucleic Acids Res*, vol. 32, p. D3–22, 2004.

[145] M. Galperin, « The molecular biology database collection : 2006 update. », *Nucleic Acids Res*, vol. 34, p. D3–5, Jan 2006.

[146] J. Gollub, C. A. Ball, G. Binkley, J. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaloper, J. C. Matese, M. Schroeder, P. O. Brown, D. Botstein et G. Sherlock, « The stanford microarray database : data access and quality assessment tools », *Nucleic Acids Res*, vol. 31, no. 1, p. 94–6, 2003.

[147] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J. C. Matese, T. Hernandez-Boussard, C. A. Rees, J. M. Cherry, D. Botstein, P. O. Brown et A. A. Alizadeh, « Source : a unified genomic resource of functional annotations, ontologies, and gene expression data », *Nucleic Acids Res*, vol. 31, no. 1, p. 219–23, 2003.

[148] J. Stoeckert, C. J., H. C. Causton et C. A. Ball, « Microarray databases : standards and ontologies », *Nat Genet*, vol. 32 Suppl, p. 469–73, 2002.

[149] R. Edgar, M. Domrachev et A. E. Lash, « Gene expression omnibus : Ncbi gene expression and hybridization array data repository », *Nucleic Acids Res*, vol. 30, no. 1, p. 207–10, 2002.

[150] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra et S. A. Sansone, « Arrayexpress–a public repository for microarray gene expression data at the ebi », *Nucleic Acids Res*, vol. 31, no. 1, p. 68–71, 2003.

[151] K. Ikeo, J. Ishi-i, T. Tamura, T. Gojobori et Y. Tateno, « Cibex : center for information biology gene expression database », *C R Biol*, vol. 326, no. 10-11, p. 1079–82, 2003.

[152] L. H. Saal, C. Troein, J. Vallon-Christersson, S. Gruvberger, A. Borg et C. Peterson, « Bioarray software environment (base) : a platform for comprehensive management and analysis of microarray data », *Genome Biol*, vol. 3, no. 8, p. SOFTWARE0003, 2002.

[153] H. Mangalam, G. Chen, J. Stewart, A. Farmer, J. Zhou, G. Colello, K. Schlauch, J. Weller et M. Waugh, « Genex : an open source gene expression database and integrated tool set », *IBM systems journal*, vol. 40, no. 2, p. 552–569, 2001.

[154] S. Arfin, A. Long, E. Ito, L. Tolleri, M. Riehle, E. Paegle et G. Hatfield, « Global gene expression profiling in escherichia coli k12. the effects of integration host factor. », *J Biol Chem*, vol. 275, p. 29672–84, Sep 2000.

[155] Y. H. Yang, M. J. Buckley, S. Dudoit et T. P. Speed, « Comparison of methods for image analysis on cdna microarray data », *Journal of Computational and Graphical Statistics*, vol. 11, p. 108–136, 2002.

[156] Y. Chen, E. R. Dougherty et M. L. Bittner, « Ratio based decisions and the quantitative analysis of cdna microarray images », *Journal of Biomedical Optics 2*, p. 364–374, 1997.

[157] « Quantarray analysis software ». `http://lifesciences.perkinelmer.com`.

[158] M. B. Eisen, « Scanalyze user manual ». `http://rana.lbl.gov`, 1999.

[159] F. Tahi, B. Achddou, C. Decraene, O. Alibert, H. Guiot, C. Auffray et G. Pietu, « Automatic quantitation of hybridization signals on cdna arrays », *Biotechniques*, vol. 32, no. 6, p. 1386–8, 1390, 1392, 1394, 1396–7, 2002.

[160] Q. Li, C. Fraley, R. E. Bumgarner, K. Y. Yeung et A. E. Raftery, « Donuts, scratches and blanks : robust model-based segmentation of microarray images », *Bioinformatics*, vol. 21, no. 12, p. 2875–82, 2005.

[161] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang et J. Zhang, « Bioconductor : open software development for computational biology and bioinformatics », *Genome Biol*, vol. 5, no. 10, p. R80, 2004.

[162] G. K. Smyth, Y. H. Yang et T. Speed, « Statistical issues in cdna microarray data analysis », *Methods Mol Biol*, vol. 224, p. 111–36, 2003.

[163] G. A. Churchill, « Fundamentals of experimental design for cdna microarrays », *Nat Genet*, vol. 32 Suppl, p. 490–5, 2002.

[164] M. L. Lee, F. C. Kuo, G. A. Whitmore et J. Sklar, « Importance of replication in microarray gene expression studies : statistical methods and evidence from repetitive cdna hybridizations », *Proc Natl Acad Sci U S A*, vol. 97, no. 18, p. 9834–9, 2000.

[165] Y. Yang et T. Speed, *Statistical analysis of gene expression microarray data*, chap. Design and analysis of comparative microarray experiments, p. 240. Chapman & HallRC, 2003.

[166] M. Black et R. Doerge, « Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. », *Bioinformatics*, vol. 18, p. 1609–16, Dec 2002.

[167] Y. Yang et T. Speed, « Design issues for cdna microarray experiments. », *Nat Rev Genet*, vol. 3, p. 579–88, Aug 2002.

[168] M. Callow, S. Dudoit, E. Gong, T. Speed et E. Rubin, « Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. », *Genome Res*, vol. 10, p. 2022–9, Dec 2000.

[169] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J. B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D. K. Slonim, C. Soderlund, W. L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. A. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, J. Hudson, J. R., S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa et N. e. a. Nomura, « A gene map of the human genome », *Science*, vol. 274, no. 5287, p. 540–6, 1996.

[170] J. Quackenbush, F. Liang, I. Holt, G. Pertea et J. Upton, « The tigr gene indices : reconstruction and representation of expressed gene sequences », *Nucleic Acids Res*, vol. 28, no. 1, p. 141–5, 2000.

[171] M. K. Kerr et G. A. Churchill, « Statistical design and the analysis of gene expression microarray data », *Genet Res*, vol. 77, no. 2, p. 123–8, 2001.

[172] Q. J., « Microarray data normalisation and transformation », *Nature Genetics Suppl.*, vol. 32, p. 496–501, 2002.

[173] J. Larkin, B. Frank, H. Gavras, R. Sultana et J. Quackenbush, « Independence and reproducibility across microarray platforms. », *Nat Methods*, vol. 2, p. 337–44, May 2005.

[174] D. Rhodes et A. Chinnaiyan, « Integrative analysis of the cancer transcriptome. », *Nat Genet*, vol. 37 Suppl, p. S31–7, Jun 2005.

[175] M. K. Kerr et G. A. Churchill, « Experimental design for gene expression microarrays », *Biostatistics*, vol. 2, no. 2, p. 183–201, 2001.

[176] M. K. Kerr, M. Martin et G. A. Churchill, « Analysis of variance for gene expression microarray data », *J Comput Biol*, vol. 7, no. 6, p. 819–37, 2000.

[177] M. L. Lee, W. Lu, G. A. Whitmore et D. Beier, « Models for microarray gene expression data », *J Biopharm Stat*, vol. 12, no. 1, p. 1–19, 2002.

[178] J. Landgrebe, F. Bretz et E. Brunner, « Efficient two-sample designs for microarray experiments with biological replications », *In Silico Biol*, vol. 4, no. 4, p. 461–70, 2004.

[179] G. K. Smyth, « Linear models and empirical bayes methods for assessing differential expression in microarray experiments », *in Statistical Applications in Genetics and Molecular Biology*, vol. 1, 2004.

[180] G. K. Smyth, « Linear models and empirical bayes methods for assessing differential expression in microarray experiments », *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. art. 3, 2004.

[181] G. Pietu, O. Alibert, V. Guichard, B. Lamy, F. Bois, E. Leroy, R. Mariage-Sampson, R. Houlgatte, P. Soularue et A. C., « Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cdna array. », *Genome Res.*, vol. 6, p. 492–503, 1996.

[182] B. Bolstad, R. Irizarry, M. Astrand et T. Speed, « A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. », *Bioinformatics*, vol. 19, p. 185–93, Jan 2003.

[183] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari et R. S. Paules, « Assessing gene significance from cdna microarray expression data via mixed models », *J Comput Biol*, vol. 8, no. 6, p. 625–37, 2001.

[184] S. Chatterjee et B. Price, *Regression Analysis by Example*. John Wiley & Sons, 1991.

[185] Y. Chen, V. Kamat, E. R. Dougherty, M. L. Bittner, P. S. Meltzer et J. M. Trent, « Ratio statistics of gene expression levels and applications to microarray data analysis », *Bioinformatics*, vol. 18, no. 9, p. 1207–15, 2002.

[186] Y. H. Yang, S. Dudoit, P. Luu et T. P. Speed, « Normalization for cdna microarray data », *Proceedings of SPIE, BIOS 2001, Microarrays :Optical Technologies and Informatics*, vol. 4266, p. 141–152, 2001.

[187] D. Finkelstein, J. Gollub, R. Ewing, F. Sterky, S. S. et J. Cherry, « Iterative linear regression by sector », *in CAMDA 2000* (S. Lin et K. A. K.F. Johnson, éds), p. 57–68, S.M. Lin and K.F. Johnson, Kluwer Academic, 2001.

[188] W. Cleveland, « Robust locally weighted regression and smoothing scatterplots », *J. Am. Stat. Assoc.*, vol. 74, p. 829–836, 1979.

[189] S. Dudoit, Y. Yang, M. Callow et T. Speed, « Statistical methods for identifying genes with differential expression in replicated cdna microarray experiments », *Stat. Sin.*, vol. 12, no. 1, p. 111–139, 2002.

[190] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai et T. P. Speed, « Normalization for cdna microarray data : a robust composite

method addressing single and multiple slide systematic variation », *Nucleic Acids Res*, vol. 30, no. 4, p. e15, 2002.

[191] G. K. Smyth et T. P. Speed, « Normalization of cdna microarray data », *METHODS*, vol. 31, no. 4, p. 265–273, 2003.

[192] I. Yang et et al, « Within the fold : assesing the differential expression measures and reproducibility in microarray assays », *Genome Biol.*, p. research0062.1–006212, 2002.

[193] J. P. Novak, R. Sladek et T. J. Hudson, « Characterization of variability in large-scale gene expression data : implications for study design », *Genomics*, vol. 79, no. 1, p. 104–13, 2002.

[194] I. Lönnste et T. Speed, « Replicated microarray data », *Satistical Sinica*, vol. 12, p. 31–46, 2002.

[195] B. Efron, R. Tibshirani, J. Storey et V. Tusher, « Empirical bayes analysis of a microarray experiment », *Journal of the American Statistical Association*, vol. 96, p. 1151–1160, 2001.

[196] V. Tusher, R. Tibshirani et G. Chu, « Significance analysis of microarrays applied to the ionizing radiation response. », *Proc Natl Acad Sci U S A*, vol. 98, p. 5116–21, Apr 2001.

[197] G. Snecdecor et W. Cochran, *Statistical methods*. Iowa States University Press, 1989.

[198] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi et I. Golani, « Controlling the false discovery rate in behavior genetics research. », *Behav Brain Res*, vol. 125, p. 279–84, Nov 2001.

[199] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts et V. Sondak, « Molecular classification of cutaneous malignant melanoma by gene expression profiling. », *Nature*, vol. 406, p. 536–40, Aug 2000.

[200] P. Legendre et L. Legendre, *Numerical Ecology*. Elsevier, 1998.

[201] A. Brazma, I. Jonassen, J. Vilo et E. Ukkonen, « Predicting gene regulatory elements in silico on a genomic scale. », *Genome Res*, vol. 8, p. 1202–15, Nov 1998.

[202] V. Filkov, S. Skiena et J. Zhi, « Analysis techniques for microarray time-series data », *in RECOMB*, p. 124–131, 2001.

[203] S. Dudoit et J. Fridlyand, « A prediction-based resampling method for estimating the number of clusters in a dataset. », *Genome Biol*, vol. 3, p. RESEARCH0036, Jun 2002.

[204] M. B. Eisen, P. T. Spellman, P. O. Brown et D. Botstein, « Cluster analysis and display of genome-wide expression patterns », *Proc Natl Acad Sci U S A*, vol. 95, no. 25, p. 14863–8, 1998.

[205] S. Hilsenbeck, W. Friedrichs, R. Schiff, P. O'Connell, R. Hansen, C. Osborne et S. Fuqua, « Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. », *J Natl Cancer Inst*, vol. 91, p. 453–9, Mar 1999.

[206] K. Yeung et W. Ruzzo, « Principal component analysis for clustering gene expression data. », *Bioinformatics*, vol. 17, p. 763–74, Sep 2001.

[207] J. Benzécri et collaborateurs, *L'analyse des données. Tome 2*, vol. Tome 2, L'analyse des correpondances. Dunod Edit., 1973.

[208] K. Fellenberg, N. Hauser, B. Brors, A. Neutzner, J. Hoheisel et M. Vingron, « Correspondence analysis applied to microarray data. », *Proc Natl Acad Sci U S A*, vol. 98, p. 10781–6, Sep 2001.

[209] W. Liebermeister, « Linear modes of gene expression determined by independent component analysis. », *Bioinformatics*, vol. 18, p. 51–60, Jan 2002.

[210] S. Lee et S. Batzoglou, « Application of independent component analysis to microarrays. », *Genome Biol*, vol. 4, no. 11, p. R76, 2003.

[211] R. R. Sokal et P. H. A. Sneath, *Principles of Numerical Taxonomy*. San Francisco : W. H. Freeman, 1963.

[212] B. Morgan et A. Ray, « Non-uniqueness and inversions in clusters analysis », *Appl. Statist.*, vol. 44, p. 117–134, 1995.

[213] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack et A. Levine, « Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. », *Proc Natl Acad Sci U S A*, vol. 96, p. 6745–6750, Jun 1999.

[214] J. Ward, « Hierarchical grouping to optimize an objective function », *J. Am. Stat. Assoc.*, vol. 58, p. 236–244, 1963.

[215] L. M. In Le Cam et J. Neyman, « Some methods for classification and analysis of multivariate observations. », *in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (U. of California Press, éd.), vol. 1, p. 281–297, 1967.

[216] L. Kaufman et P. J. Rousseeuw, *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley-Interscience Publication, 1990.

[217] M. Kerr et G. Churchill, « Bootstrapping cluster analysis : assessing the reliability of conclusions from microarray experiments. », *Proc Natl Acad Sci U S A*, vol. 98, p. 8961–5, Jul 2001.

[218] T. Kohonen, *Self-organizing maps*. Springer, 1995.

[219] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander et T. R. Golub, « Interpreting patterns of gene expression with self-organizing maps : methods and application to hematopoietic differentiation », *Proc Natl Acad Sci U S A*, vol. 96, no. 6, p. 2907–12, 1999.

[220] P. Törönen, M. Kolehmainen, G. Wong et E. Castrén, « Analysis of gene expression data using self-organizing maps. », *FEBS Lett*, vol. 451, p. 142–6, May 1999.

[221] D. Dembélé et P. Kastner, « Fuzzy c-means method for clustering microarray data. », *Bioinformatics*, vol. 19, p. 973–80, May 2003.

[222] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer et Z. Yakhini, « Tissue classification with gene expression profiles. », *J Comput Biol*, vol. 7, no. 3-4, p. 559–83, 2000.

[223] S. Dudoit et J. Fridlyand, « Bagging to improve the accuracy of a clustering procedure. », *Bioinformatics*, vol. 19, p. 1090–9, Jun 2003.

[224] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[225] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, J. Ares, M. et D. Haussler, « Knowledge-based analysis of microarray gene expression data by using support vector machines », *Proc Natl Acad Sci U S A*, vol. 97, no. 1, p. 262–7, 2000.

[226] N. Iizuka, M. Oka, H. Yamada-Okabe, M. Nishida, Y. Maeda, N. Mori, T. Takao, T. Tamesa, A. Tangoku, H. Tabuchi, K. Hamada, H. Nakayama, H. Ishitsuka, T. Miyamoto, A. Hirabayashi, S. Uchimura et Y. Hamamoto, « Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. », *Lancet*, vol. 361, p. 923–9, Mar 2003.

[227] J. R. Quinlan, *Machine Learning*, chap. Induction of decision trees, p. 81–106. 1986.

[228] A. Boulesteix, G. Tutz et K. Strimmer, « A cart-based approach to discover emerging patterns in microarray data. », *Bioinformatics*, vol. 19, p. 2465–72, Dec 2003.

[229] R. C. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry et S. Dudoit, *Bionformatics and Computational Biology Solution Using R and Bioconductor*. Statistics for Biology and Health, Springer, 2005.

[230] C. Nguyen, D. Rocha, S. Granjeaud, M. Baldit, K. Bernard, P. Naquet et B. Jordan, « Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cdna clones. », *Genomics*, vol. 29, p. 207–16, Sep 1995.

[231] G. Piétu, O. Alibert, V. Guichard, B. Lamy, F. Bois, E. Leroy, R. Mariage-Sampson, R. Houlgatte, P. Soularue et C. Auffray, « Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cdna array. », *Genome Res*, vol. 6, p. 492–503, Jun 1996.

[232] F. Bertucci, K. Bernard, B. Loriod, Y. Chang, S. Granjeaud, D. Birnbaum, C. Nguyen, K. Peck et B. Jordan, « Sensitivity issues in dna array-based expression measurements and performance of nylon microarrays for small samples. », *Hum Mol Genet*, vol. 8, p. 1715–22, Sep 1999.

[233] R. Shields, « Miame, we have a problem. », *Trends Genet*, vol. 22, p. 65–6, Feb 2006.

[234] R Development Core Team, *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[235] J. Welsh, L. Sapinoso, S. Kern, D. Brown, T. Liu, A. Bauskin, R. Ward, N. Hawkins, D. Quinn, P. Russell, R. Sutherland, S. Breit, C. Moskaluk, H. Frierson et G. Hampton, « Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. », *Proc Natl Acad Sci U S A*, vol. 100, p. 3410–5, Mar 2003.

[236] C. Coulouarn, G. Lefebvre, C. Derambure, T. Lequerre, M. Scotte, A. Francois, D. Cellier, M. Daveau et J. Salier, « Altered gene expression in acute systemic inflammation detected by complete coverage of the human liver transcriptome. », *Hepatology*, vol. 39, p. 353–64, Feb 2004.

[237] C. Gangneux, M. Daveau, M. Hiron, C. Derambure, J. Papaconstantinou et J. Salier, « The inflammation-induced down-regulation of plasma fetuin-a (alpha2hs-glycoprotein) in liver results from the loss of interaction between long c/ebp isoforms at two neighbouring binding sites. », *Nucleic Acids Res*, vol. 31, p. 5957–70, Oct 2003.

[238] D. Odom, N. Zizlsperger, D. Gordon, G. Bell, N. Rinaldi, H. Murray, T. Volkert, J. Schreiber, P. Rolfe, D. Gifford, E. Fraenkel, G. Bell et R. Young, « Control of pancreas and liver gene expression by hnf transcription factors. », *Science*, vol. 303, p. 1378–81, Feb 2004.

[239] H. Inoue, W. Ogawa, M. Ozaki, S. Haga, M. Matsumoto, K. Furukawa, N. Hashimoto, Y. Kido, T. Mori, H. Sakaue, K. Teshigawara, S. Jin, H. Iguchi, R. Hiramatsu, D. LeRoith, K. Takeda, S. Akira et M. Kasuga, « Role of stat-3 in regulation of hepatic gluconeogenic genes and carbohydrate metabolism in vivo. », *Nat Med*, vol. 10, p. 168–74, Feb 2004.

[240] J. García-Martínez, A. Aranda et J. Pérez-Ortín, « Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. », *Mol Cell*, vol. 15, p. 303–13, Jul 2004.

[241] T. Kawai, J. Fan, K. Mazan-Mamczarz et M. Gorospe, « Global mrna stabilization preferentially linked to translational repression during the endoplasmic reticulum stress response. », *Mol Cell Biol*, vol. 24, p. 6773–87, Aug 2004.

[242] G. Laroia, B. Sarkar et R. Schneider, « Ubiquitin-dependent mechanism regulates rapid turnover of au-rich cytokine mrnas. », *Proc Natl Acad Sci U S A*, vol. 99, p. 1842–6, Feb 2002.

[243] R. Winzen, G. Gowrishankar, F. Bollig, N. Redich, K. Resch et H. Holtmann, « Distinct domains of au-rich elements exert different functions in mrna destabilization and stabilization by p38 mitogen-activated protein kinase or hur. », *Mol Cell Biol*, vol. 24, p. 4835–47, Jun 2004.

[244] T. Bakheet, B. Williams et K. Khabar, « Ared 2.0 : an update of au-rich element mrna database. », *Nucleic Acids Res*, vol. 31, p. 421–3, Jan 2003.

[245] C. Derambure, C. Coulouarn, F. Caillot, G. Lefebvre, M. Hiron, M. Scotte, A. Francois, C. Duclos, O. Goria, M. Gueudin, C. Cavard, B. Terris, M. Daveau et J. P. Salier, « Genome-wide differences in hepatitis c- vs alcoholism-associated hepatocellular carcinoma », *Hepatology*, vol. soumis, 2006.

[246] C. Coulouarn, C. Derambure, G. Lefebvre, R. Daveau, M. Hiron, M. Scotte, A. François, M. Daveau et J. Salier, « Global gene repression in hepatocellular carcinoma and fetal liver, and suppression of dudulin-2 mrna as a possible marker for the cirrhosis-to-tumor transition. », *J Hepatol*, vol. 42, p. 860–9, Jun 2005.

[247] C. Coulouarn, G. Lefebvre, R. Daveau, F. Letellier, M. Hiron, L. Drouot, M. Daveau et J. Salier, « Genome-wide response of the human hep3b hepatoma cell to proinflammatory cytokines, from transcription to translation. », *Hepatology*, vol. 42, p. 946–55, Oct 2005.

Résumé

La phase aiguë de l'inflammation sytémique est une période de transition subséquente au trauma, coordonnée par des médiateurs telles les cytokines. Ces dernières ciblent majoritairement le foie, lequel régule alors principalement la production des APPs et des APRIPs.
Ce travail montre d'abord la création d'une plate-forme à puces à ADN dédiée à l'étude du transriptome hépatique humain et fondée d'une part sur le développement de la puce elle-même par la désignation des sondes ADNc qui la définissent, d'autre part sur la création d'une base de données standardisée et également sur l'écritre de programmes d'analyse statistique.
Il montre enfin comment cette plate-forme a permis de mettre en évidence d'une part une corrélation entre 134 transcripts et la degré de l'inflammation et d'autre part une liste de gènes répartis en 12 groupes fonctionnels subissant une altération de leur transcription, de leur stabilité post-transcriptionnelle et de leur traduction durant la phase aiguë de linflammation.

Mots clefs

analyse statistique, base de donnée, classification, cytokines, foie, gènes à expression différentielle, inflammation sytémique aiguë, puces à ADN, regroupement, stabilité des ARNm, sélection de sondes, tissus, transcriptome

Summary

The inflammation acute phase is a transitional period that follows the trauma, predominantly coordinated by mediators such as cytokines. The latter mainly targets the liver which then regulates among others APP and APRIPS production.
This work initially shows the creation of a microarray platform dedicated to human hepatic transcriptome analysis and founded on the one hand on the development of the microarray itself by indicating cDNA probes which define it, on the other hand on the creation of a standardized database and also by the writing of statistical analysis software.
It shows finally how this platform allowed to highlight on the one hand a correlation between both 134 transcripts and inflammation degree and on the other hand a list of genes clustered in 12 functional groups undergoing modifications of their transcription, their post-transcriptionnal stability and their translation during the inflammation acute phase.

Keywords

acute sytemic inflammation, classification, clustering, cytokines, differential expressed genes, microarray, database, liver, mRNA decay, probe selection, tissue, statistical analysis, transcriptome