Modélisation, classification et fusion de données biomédicales

Vincent Barra

Ecole Doctorale
des Sciences pour l'Ingénieur

N° d'ordre :

Habilitation à Diriger des Recherches

Présentée le 29 avril 2004 à l'Université Blaise Pascal

par

Vincent BARRA

MODELISATION, CLASSIFICATION ET FUSION DE DONNEES BIOMEDICALES

Applications à l'imagerie du cerveau et des puces à ADN

Rapporteurs

Jacques DEMONGEOT	Professeur à l'Université Grenoble I
Serge HAZOUT	Professeur à l'Université Paris 7
Alain QUILLIOT	Professeur à l'ISIMA

Examinateurs

Jean-Yves BOIRE	Maître de conférences, praticien hospitalier à l'Université d'Auvergne
Jean-Marc PETIT	Maître de conférences, HDR à l'Université Blaise Pascal

.

AVANT PROPOS

Ce mémoire rapporte quatre années de travail, autour duquel ont gravité un ensemble de personnes que je désire remercier ici.

Je souhaite tout d'abord exprimer toute ma gratitude à Jacques DEMONGEOT, professeur à l'Université Joseph Fourier de Grenoble et directeur du TIMC-IMAG, qui a accepté de rapporter ce mémoire après avoir évalué mon travail de thèse. Un grand merci également à Serge HAZOUT, professeur à l'Université Denis Diderot de Paris et directeur de l'équipe INSERM E0346, pour avoir accepté de rapporter ce travail.

A cheval sur deux laboratoires, mes travaux ont été suivis et encadrés par Jean-Yves BOIRE et Alain QUILLIOT, qui m'ont fait pleinement confiance et m'ont donné les moyens de progresser et de réussir. Qu'ils en soient sincèrement remerciés ici.

Plus près de moi, je souhaiterais exprimer toute mon amitié à David "Benny" HILL, qui m'a accueilli en 2001 au sein de la thématique bioinformatique du LIMOS, et à Jean-Marc PETIT, pour avoir accepté de participer au jury de cette habilitation. Merci également à Pierre PEYRET pour sa grande disponibilité, son enthousiasme et ses précieuses remarques sur la biologie des puces à ADN.

Ayant passé une bonne partie de ces quatre dernières années dans mes lieux d'enseignement, je souhaiterais associer aux remerciements l'ensemble du personnel du département imagerie de l'IUT d'Aubière, au Puy en Velay, et la quasi totalité des collègues de Vichy, avec une mention spéciale pour Marie, qui me supporte dans le bureau depuis bientôt 2 ans, et Véro, pour tous les intéressants moments que nous passons ensemble.

Enfin, un travail de ce type ne saurait être possible sans la compagnie et l'appui de camarades et d'amis. Merci donc à Alice, Christophe, Emmanuelle, Julien, Laurent, Magali, Sébastien(s) et aux p'tits gars de l'INSERM pour les apéros, les repas (ah, le bourbonnais), les cours de biologie et autres bons moments passés ensemble.

SOMMAIRE

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

PREAMBULE

Ce mémoire synthétise les travaux que j'ai menés de 2000 à 2004, au sein de deux laboratoires des facultés de Clermont-Ferrand : l'Equipe de Recherche en Imagerie Médicale (ERIM, Université d'Auvergne), où j'ai effectué ma thèse, et le Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS, Université Blaise Pascal) dans lequel j'ai été accueilli suite à mon recrutement en tant que maître de conférences dans cette même université.

Ce changement de laboratoire s'est accompagné d'une modification de mon thème principal de recherche, passant du traitement d'images médicales multimodales par des techniques de fusion d'informations, au domaine de la bioinformatique en général, et de l'étude des puces à ADN en particulier.

Plutôt que d'essayer de regrouper artificiellement ces deux thèmes au sein d'un même plan, j'ai préféré diviser ce mémoire en deux parties distinctes et cohérentes, chacune traitant d'un des deux aspects de recherche que je mène actuellement de front.

Ainsi, la première partie résume les travaux que j'ai effectués depuis 2001 dans le domaine de la fusion de données appliquée au traitement d'images 3D du cerveau, soit directement soit dans le cadre du co-encadrement de deux doctorants. Le dernier chapitre de cette partie met en particulier en perspective les nouveaux développements espérés sur la stimulation magnétique transcrânienne, à travers l'encadrement d'une thèse CIFRE que j'assure par délégation à temps plein.

La seconde partie se concentre sur les recherches que je mène depuis septembre 2001 au LIMOS, concernant l'étude des images de puces à ADN. J'expose dans cette partie au travers de trois chapitres mon projet de recherche dans ce domaine, et je présente pour chaque choix retenu ma contribution sous la forme d'un simulateur d'images de biopuces transcriptome et de nouvelles méthodes d'analyse de ces images.

Si les deux parties sont clairement décorrélées, j'ai néanmoins essayé de dégager une problématique générale commune à mes travaux, que j'ai nommée sans forfanterie modélisation, classification et fusion de données biomédicales, et qui constitue le titre de ce manuscrit.

PREMIERE PARTIE :
FUSION D'IMAGES 3D DU CERVEAU

Introduction

Le secret d'un bon discours, c'est d'avoir une bonne introduction
et une bonne conclusion. Ensuite, il faut s'arranger pour ces deux parties
ne soient pas très éloignées l'une de l'autre.
George Burns.

Avec le développement des dossiers médicaux informatiques et la généralisation des techniques d'imagerie, il devient possible, pour une pathologie donnée, de disposer d'un grand nombre de données hétérogènes, complémentaires et parfois ambiguës. Le clinicien, analysant ces multiples informations, opère une agrégation de celles-ci, en fonction de jugements subjectifs et approximatifs fondés sur sa propre expérience. Le but de ce raisonnement est de synthétiser un état de la pathologie le plus complet possible, par exemple pour proposer un diagnostic, établir un pronostic ou même élaborer une aide à l'intervention chirurgicale.

Ces dernières années, des modélisations formelles de cette attitude ont été construites, fondées pour la plupart sur des approches prenant en compte les redondances, les complémentarités et les ambiguïtés inhérentes aux données médicales. Regroupées sous l'appellation "fusion", ces modèles ont pour but de gérer au mieux ces différents aspects pour faire converger les connaissances et proposer une information synthétique la plus exploitable possible.
Nous avons proposé dès 1997 au sein de l'Equipe de Recherche en Imagerie Médicale (ERIM, Université d'Auvergne, Clermont-Ferrand), d'appliquer ces concepts de fusion de données à plusieurs études concernant le cerveau. Le problème principal qui survient lorsqu'il s'agit de traiter des données médicales est que la "vérité" concernant les informations n'existe pas ou n'est pas nécessairement accessible. D'une manière plus générale, l'information exprimée n'est qu'une approximation, une théorisation ou une représentation de l'objet ou du phénomène qui est mesuré. De plus, ces informations sont généralement en grand nombre et l'agrégation de telles données devient de ce fait en général complexe. Cette complexité s'exprime enfin également à travers la nature complémentaire et redondante des informations, nature qu'il convient de prendre en compte pour exprimer au mieux par exemple l'état d'une zone pathologique (complémentarité, conflit et ambiguïté) au milieu de régions saines (redondances).

Nous avons alors construit une méthodologie fondée sur l'agrégation de l'ensemble des connaissances disponibles dans une étude donnée, modélisées dans un cadre théorique commun permettant de gérer les imprécisions et les incertitude inhérentes aux données du vivant. Si de nombreuses méthodes de modélisation sont disponibles dans ce domaine, le contexte théorique retenu a été celui de la logique possibiliste, en raison notamment de la richesse des étapes de modélisation et de fusion des données proposée.

Cette première partie propose un état de l'avancement de nos recherches dans ce domaine. Les fondements de la méthode et les premières applications cliniques sont tout d'abord rappelés dans le premier chapitre, et les chapitres suivants décrivent les développements réalisés depuis 2001 dans le cadre de mes propres travaux ou de ceux de deux doctorants que j'ai été amené à co-encadrer. Ainsi, la segmentation de structures cérébrales est tout d'abord abordée dans le chapitre 2. L'intérêt d'une telle

segmentation est tout d'abord démontré, puis une nouvelle méthode fondée sur les concepts de fusion d'informations est proposée, validée et appliquée à l'étude de pathologies diverses. Le chapitre 3 décrit quant à lui deux approches de quantification d'activités fonctionnelles cérébrales, la première fondée sur une segmentation préalable de structures cérébrales d'intérêt, et la seconde sur une approche novatrice prenant en compte le caractère discret des pavages de voxels dans les images. Pour ces deux applications, les méthodes proposées et les problématiques cliniques envisagées sont décrites, et nombre de perspectives, à la fois fondamentales et cliniques, sont exposées. Enfin, le chapitre 4 est prospectif et concerne une introduction à l'étude de la stimulation magnétique transcrânienne, débutée en septembre 2003 dans le cadre d'une thèse CIFRE que j'encadre à temps plein. Y sont exposées les bases des développements futurs que nous souhaitons réaliser dans ce domaine, à la fois en relation avec les concepts de fusion qui sont fortement intégrés à l'ERIM, et en proposant de nouvelles approches.

Chapitre 1 - Rappel des épisodes précédents

Se rappeler quelque chose est encore le meilleur moyen de ne pas l'oublier.
Pierre Dac.

Introduction
Le recueil de données diverses, issues tant de l'imagerie que de connaissances expertes ou de signaux physiologiques, est devenu courant dans les services cliniques pour l'étude d'une pathologie donnée. L'exploitation de l'ensemble de ces renseignements, effectuée par le clinicien qui analyse et agrège les données en fonction de ses connaissances, conduit généralement à un diagnostic plus précis, plus clair et plus fiable. La principale motivation du travail entrepris de 1997 à 2000 dans l'Equipe de Recherche en Imagerie Médicale reposait sur une modélisation de ce raisonnement fondée sur un processus de fusion d'informations. Ce chapitre rappelle le contexte théorique introduit, celui de la logique possibiliste, les méthodes originales de modélisation d'informations, de fusion et de décision développées et les premières applications envisagées et validées cliniquement de 1997 à 2000. Ces dernières concernent d'une part l'agrégation d'images morphologiques, d'autre part la fusion d'images anatomiques et fonctionnelles, et enfin l'agrégation d'images et de connaissances expertes formalisées.

1. Cadre théorique

1.1. Eléments de la théorie des possibilités

Le cadre théorique retenu est celui de la théorie des possibilités, introduite en 1978 par Zadeh [ZADEH78], puis développée par Dubois et Prade [DUBOIS88]. Ce choix a été motivé par un certain nombre de facteurs, depuis la souplesse dans la modélisation des informations jusqu'à l'interprétation des imprécisions et incertitudes de nature non probabiliste.
Cette théorie peut être vue indépendamment de toute interprétation probabiliste comme une approche ordinale de l'incertain dans [0,1], exploitée à l'aide des mesures de possibilité.

1.1.1. Mesure et distribution de possibilité

Soit X un ensemble de référence fini. Une mesure de possibilité Π est définie sur l'ensemble $P(X)$ des parties de X et prend ses valeurs dans [0,1] telle que :

(i) $\Pi(\varnothing)=0$, $\Pi(X)=1$

(ii) $(\forall i \in N)\ (\forall Ai \in P(X))$ $\quad \Pi\left(\bigcup_i A_i\right) = \sup_i \Pi(A_i)$

La réalisation de l'un des événements A_i pris indifféremment est donc affectée du même coefficient de possibilité que la réalisation de l'événement le plus possible. Si Π permet de déterminer le degré avec lequel l'union d'événements dont on sait à quel point ils sont possibles, sera elle-même un événement possible, elle ne permet pas de se prononcer sur l'intersection d'événements. Les conditions (i) et (ii) imposent seulement que le coefficient attribué à l'intersection des événements soit majoré par le plus petit des coefficients attribués à chacun d'entre eux.
Une mesure de possibilité Π est totalement définie par la donnée de $2^{|X|}$ coefficients de possibilité attribués à chaque élément de $P(X)$. Par (ii), il est également possible de définir Π en indiquant uniquement les coefficients attribués aux singletons de X. Ceci introduit la notion de distribution de possibilité par :

$$\pi : X \to [0,1]$$
$$x \mapsto \Pi(\{x\})$$

et contrainte par :

$$\underset{x \in X}{Sup}\, \pi(x) = 1$$

Une distribution de possibilité peut être reliée à un sous-ensemble flou par l'intermédiaire de la fonction d'appartenance à cet ensemble.

1.1.2. Mesure de nécessité

Une mesure de possibilité fournit une information sur l'occurrence d'un événement A relatif à un ensemble de référence X, mais elle ne suffit pas pour décrire l'incertitude existante sur cet événement. Ainsi, si $\Pi(A)=1$ et $\Pi(A^c)=1$, la réalisation de A et complètement indéterminée, alors qu'elle est certaine si $\Pi(A^c)=0$. Pour compléter l'information sur A, on indique donc le degré avec lequel la réalisation de A est certaine par l'intermédiaire d'une mesure de nécessité $N : P(X) \to [0,1]$ définie par :

(i) $N(\varnothing)=0,\ N(X)=1$

(ii) $(\forall i \in N)\ (\forall A_i \in P(X))\ N\left(\bigcap_i A_i\right) = \underset{i}{Inf}\ N(A_i)$

Plus un événement A est affecté d'une grande nécessité, moins l'événement complémentaire A^c est possible, donc plus on est certain de la réalisation de A.

1.1.3. Variables linguistiques et propositions floues

Zadeh a introduit la théorie des possibilités à propos de la caractérisation de variables par des descriptions linguistiques imprécises, représentées par des sous-ensembles flous. La fonction d'appartenance de ceux-ci conduit à la définition d'une distribution de possibilité qui permet de traiter les incertitudes engendrées au cours d'un raisonnement fondé sur les caractéristiques floues des variables.

Un certain nombre de connaissances utilisées dans la suite sont issues de discussions avec un expert. Ces connaissances sont décomposées en assertions élémentaires, chacune étant modélisée à l'aide de variables linguistiques.

Une variable linguistique est représentée par un triplet (V, X, T_V), où V est une variable définie sur un ensemble de référence X, pouvant prendre une valeur quelconque de X. $T_V=\{A_i\}$ est un ensemble de sous-ensembles flous de X utilisés pour décrire V. Le but de l'utilisation des variables linguistiques caractérisées par des descriptions floues est d'éviter les bornes artificiellement rigides des descriptions ordinaires et d'introduire une grande souplesse dans les caractérisations. Ces dernières sont soit des éléments de T_v, soit des formes modifiées de ces éléments par des adverbes du type «très», «plus», «moins», appelés modificateurs linguistiques.

Un modificateur linguistique est décrit par un opérateur m qui permet, à partir de toute caractérisation floue $A \in T_V$ de produire une nouvelle caractérisation floue $m(A)$. Si la fonction d'appartenance à l'ensemble flou A est f_A, celle de $m(A)$ est obtenue par l'intermédiaire d'une transformation attachée à m et appliquée à f_A. Pour un ensemble M de modificateurs disponibles, $M(T_V)= M \times T_V$ représente l'ensemble des descriptions possibles de V.

La représentation puis le traitement d'informations de type symbolique (typiquement une description orale donnée par un expert) passe donc par l'utilisation de variables linguistiques. Soit L un ensemble de variables linguistiques et M un ensemble de modificateurs. Une proposition floue élémentaire est définie à partir d'une variable linguistique de L par la qualification «V est A», pour une caractérisation floue A, appartenant à T_v ou à $M(T_v)$. La proposition «V est A» est d'autant moins vraie que la valeur exacte de V, soit l'élément x de X, satisfait mal la caractérisation A, c'est-à-dire que $f_A(x)$ est faible.

Une proposition floue générale est obtenue par la composition logique de propositions floues élémentaires, et induit une distribution de possibilité $\pi_{V,A}$ sur X, définie à partir de la fonction d'appartenance associée à A par :

$$(\forall x \in X)\ \pi_{V,A}(x)=f_A(x)$$

Si ε est le degré d'appartenance d'un élément $x \in X$ à la caractérisation floue A, la possibilité pour que la variable V prenne la valeur x sachant que V est caractérisée par A est égale à ε.

1.2. Fusion d'informations dans le cadre possibiliste

J'ai défini dans [BARRA00-A] la fusion d'informations comme une agrégation d'informations ambiguës, conflictuelles, complémentaires et redondantes, autorisant une interprétation des données plus précise et/ou moins incertaine. J'ai alors envisagé cette agrégation sous l'angle possibiliste, la théorie sous-jacente apportant une large gamme d'opérateurs de fusion des différentes informations.

1.2.1. Classification des opérateurs de fusion

L'ensemble des informations disponibles est modélisé par autant de distributions de possibilité π_i. L'étape de fusion créée une distribution fusionnée $\pi = F(\pi_i)$, prenant en compte l'ambiguïté et la complémentarité entre les données, par l'intermédiaire d'un opérateur de fusion F. L'idée générale derrière une approche possibiliste de la fusion d'informations est qu'il n'existe pas de mode unique de combinaison. Tout dépend de la situation étudiée et de la confiance accordée aux différentes sources d'information. Entre les comportements extrêmes conjonctif (qui exploite l'information commune aux mesures) et disjonctif (qui augmente la certitude sur l'événement observé et exprime la redondance entre les informations), il existe de nombreux opérateurs (familles paramétrées, formes explicites) dont les comportements vis-à-vis des données d'entrée peuvent être [BLOCH96] :

- constants et indépendants du contexte (CCIC) ;
- variables et indépendants du contexte (CVIC) ;
- dépendants du contexte (CDC).

Une liste non exhaustive de tels opérateurs est alors proposée et étudiée par Bloch [BLOCH96], et le Tableau 1-1 en présente un extrait.

	Sévère		Prudent		Indulgent	
CCIC	$\max(0,x+y-1)$	xy	$\min(x,y)$ ← $a=0$	$\sqrt{xy}$ $(x+y)/2$ $\text{med}(x,y,a)$	$\max(x,y)$ → $a=1$	$x+y-xy$ $\min(1,x+y)$
CVIC	$\max(0,(x^p+y^p-1)^{1/p})$ ← $a=1$		$i(x,y)^a u(x,y)^{1-a}$			$a=0$ →
	i : T-norme ; u : T-conorme					
CDC	Sources en accord		Sources en conflit partiel		Sources en désaccord	

Tableau 1-1 : quelques opérateurs en théorie des possibilités rangés suivant un ordre partiel

1.2.2. Décision

La règle de décision en théorie des possibilités est celle du maximum de possibilité. Chaque point de mesure se voit affecté à l'hypothèse pour laquelle il a le plus grand degré de possibilité. Des contraintes peuvent être ajoutées à cette règle pour modifier son comportement (test de la validité de la décision par rapport à la valeur absolue du maximum de possibilité, test du pouvoir discriminatoire de la fusion en comparant les deux valeurs les plus grandes,…).

Ayant choisi ce cadre de représentation des informations, j'ai proposé dans mon travail de thèse divers outils et méthodes de traitement de données relatives à l'imagerie cérébrale, les principaux étant rappelés dans le paragraphe suivant.

2. Outils développés

2.1. Segmentation de tissus cérébraux

En supposant que les informations pertinentes sont portées sur une image morphologique, typiquement une IRM, par la distribution des tissus cérébraux (les tissus concernés dans la suite sont la matière blanche –MB, la matière gris –MG et le liquide cérébro-spinal – LCS), j'ai tout d'abord développé une méthode permettant de segmenter ces tissus de façon fiable, reproductible et si possible indépendante du protocole d'acquisition (la séquence utilisée en IRM par exemple). L'algorithme résultant [BARRA98-B][BARRA00-B] est une méthode de classification possibiliste sur des vecteurs de coefficients d'ondelettes, donnant pour chaque voxel un ensemble de degrés d'appartenance à chaque classe de tissu T, par l'intermédiaire de distributions de possibilité π_T représentées par des cartes floues d'appartenance (Figure 1-1). Le niveau de gris de chaque voxel v dans la carte d'appartenance au tissu T est donné par la valeur $\pi_T(v)$.

Coupe IRM Carte MB Carte MG Carte LCS

Figure 1-1 : exemple de résultat de l'algorithme de caractérisation tissulaire

2.2. Modèles de processus de fusion

La fusion d'informations précédemment définie peut se décomposer en trois grandes étapes :

i. modélisation des informations dans un cadre théorique commun ;
ii. fusion des informations issues de la modélisation précédente ;
iii. prise de décision.

Le paragraphe précédent a introduit la modélisation de l'information dans le cas où cette dernière est portée par les distributions des tissus. La prise en compte des informations linguistiques et symboliques quant à elle a été introduite dans le paragraphe 1.1.3, et sera plus spécifiquement développée dans le chapitre suivant.

Mes travaux se sont alors travaux plus particulièrement concentrés sur les étapes (ii) et (iii) du processus, dans le cadre de :

- la fusion d'images IRM ;
- la fusion d'images anatomiques et fonctionnelles ;
- la fusion d'une image IRM et d'informations symboliques.

Pour chaque application, j'ai proposé un modèle initial de fusion d'informations, en terme d'opérateur d'agrégation et de règle de décision permettant d'exploiter les résultats du processus, et ai d'autre part démontré les intérêts industriels de ce type d'approche [BARRA00-H][BARRA01-C].

2.2.1. Fusion d'images anatomiques

La première fusion envisagée concerne l'agrégation de N images IRM issues de différentes techniques d'acquisition. Les informations à combiner sont homogènes, et suivant le type d'acquisition l'image offrira des contrastes plus ou moins prononcés entre les tissus ou entre parenchyme et pathologie. Un des intérêts principaux de la fusion est alors d'exploiter la complémentarité entre les différentes images pour par exemple détecter des tissus tumoraux, ou quantifier les volumes de tissus.

Dans ce cadre, j'ai proposé [BARRA00-C][BARRA00-D] et validé [BARRA99][BARRA00-E] un schéma de fusion reprenant la segmentation des tissus par classification possibiliste, et agrégeant pour un tissu T

donné les N distributions de possibilité résultantes à l'aide d'un opérateur dépendant du contexte et soulignant les redondances entre images. Plus précisément, les degrés d'appartenance $\{\pi^i_T(v)\}_{1\le i\le N}$ de chaque voxel v au tissu T calculés sur les N images sont agrégés en un degré d'appartenance $\pi_T(v)$ par :

$$\pi_T(v) = max\left(\frac{\min_i\left(\pi^i_T(v)\right)}{h}, min\left(\max_i\left(\pi^i_T(v)\right), 1-h\right)\right)$$

où h est une mesure d'accord entre les N distributions de possibilité (par exemple la distance moyenne entre les N cartes d'appartenance au tissu T).
La décision, enfin, est du type segmentation et consiste à affecter à chaque voxel v le tissu pour lequel il a le plus grand degré d'appartenance :

$$(\forall v)(v\in T) \Leftrightarrow ((\forall S\neq T)\ \pi_T(v)\geq\pi_S(v))$$

L'ensemble du processus est illustré dans le cas N=2 sur la Figure 1-2.

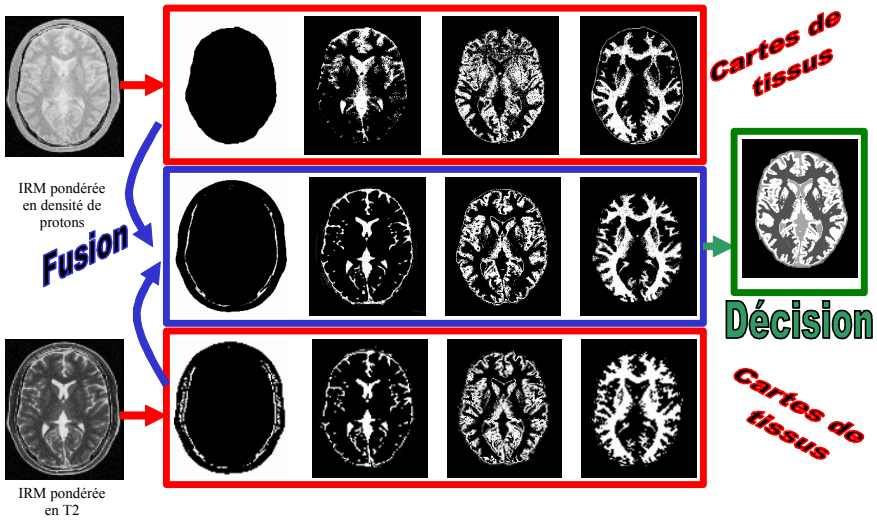


Figure 1-2 : processus de fusion de deux images IRM

2.2.2. Fusion d'images anatomiques et fonctionnelles

Le second type de fusion envisagé concerne l'agrégation de N images anatomiques et fonctionnelles (tomoscintigraphie TEM, tomographie par émission de positons TEP). Les informations à combiner sont cette fois-ci hétérogènes et les images ne sont pas informatives sur ce qu'elles ne sont pas censées représenter. Il est donc important de maîtriser la signification des niveaux de gris des images (fonctionnelles en particulier) pour une exploitation judicieuse des résultats. Les applications potentielles vont par exemple de l'étude de pathologies neurodégénératives à l'exploitation de l'imagerie de neurotransmission.
J'ai proposé pour cette approche un cadre formel d'agrégation [BARRA01-D], modélisant les données issues des images à l'aide de l'algorithme de classification possibiliste déjà présenté, et agrégeant les N distributions de possibilité d'un tissu T donné à l'aide d'un opérateur fondé sur la théorie de l'information. Deux images de décision sont enfin envisagées dans ce type d'agrégation, l'une étiquetant les voxels par la règle du maximum de possibilité, l'autre synthétisant l'ensemble des informations disponibles (morphologiques et fonctionnelles) en une seule et même image. Cette dernière fait l'objet du chapitre 3.

2.2.3. Fusion d'images et d'informations symboliques

J'ai enfin proposé en 2001 [BARRA01-A] les prémisses d'une méthode d'agrégation d'informations issues à la fois d'images et de données expertes fournies par un clinicien. Ce travail a fait l'objet de développements lors de la thèse d'Emmanuelle Frenoux , que j'ai co-encadré, et sera donc plus spécifiquement détaillée dans le chapitre 2.

3. Premières applications cliniques

Les premières applications cliniques on été essentiellement axées sur la fusion d'images anatomiques et fonctionnelles, dans le cadre de l'étude de pathologies précises.

3.1. Etude de la démence de type Alzheimer

La démence de type Alzheimer (DTA) est la première affection neurodégénérative en France en fréquence de recrutement. Elle touche uniquement le cerveau et provoque une dégénérescence des neurones, en particulier ceux impliqués dans la mémoire et les fonctions intellectuelles.
Nous disposions pour cette étude [BARRA00-G] de 9 couples d'images (IRM, TEM) recalées. La première image (séquence Flash3D) donne accès à l'information morphologique du patient, la seconde (acquise en utilisant $^{99}Tc^{m}$-HMPAO comme traceur de perfusion) l'information fonctionnelle, en précisant en particulier les réductions du métabolisme et de la perfusion, surtout sensibles pour cette pathologie dans les zones pariétales et temporales du cortex.

3.1.1. Images de fusion

Le processus de fusion a permis d'agréger les informations complémentaires issues des deux sources, et de proposer de nouvelles images synthétisant l'ensemble des données disponibles. En particulier, la génération d'une image de synthèse, développée en détail dans le chapitre 3, a donné lieu à une validation multicentrique, et fait l'objet depuis 2002 d'une valorisation dans le cadre du contrat RNTS (Réseau National Technologies de la Santé) « FusPark ». La Figure 1-3 propose à titre d'illustration une coupe de l'image de synthèse réalisée sur un des patients de l'étude, pour lequel une hypoperfusion dans les régions temporo-occipito-pariétale et temporo-occipitale a été diagnostiquée en TEM.

Image IRM Image TEM Image de synthèse

Figure 1-3 : image de synthèse issue de la fusion IRM/TEM

3.1.2. Création de cartes fusionnées pertinentes

Toute la souplesse du processus de fusion a été exploitée pour fournir au clinicien le maximum d'informations.
J'ai tout d'abord proposé [BARRA01-D] de créer une classe pathologique « matière grise hypofixante », puisque la DTA se caractérise entre autres par une perte neuronale au niveau cortical. Les zones d'intérêt étant constituées de voxels appartenant anatomiquement à la matière grise mais ayant une mesure de perfusion non conforme à celle de ce tissu dans l'image TEM, j'ai proposé de combiner les distributions de possibilité π_{IRM}^{MG} et $\pi_{TEM}^{\overline{MG}}$ à l'aide d'un simple opérateur minimum. La distribution résultante, représentant les zones hypofixantes, a été validée d'abord sur des données de synthèse (Figure 1-4), puis *a posteriori* sur les cas cliniques étudiés.

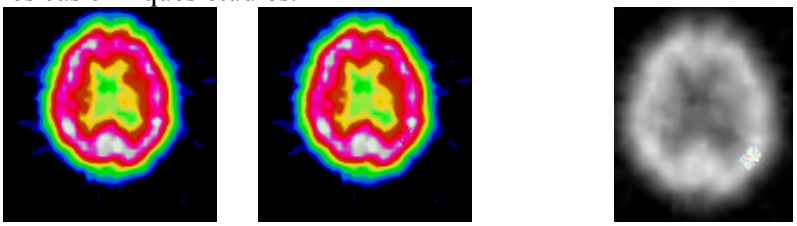


Image TEM Hypofixation simulée Détection des zones hypofixées

Figure 1-4 : classe matière grise hypofixante

Profitant de l'ensemble des informations disponibles, j'ai d'autre part proposé un modèle permettant de corriger les effets de volume partiel observés sur l'image TEM, en combinant les voxels qui appartiennent

anatomiquement au LCS, mais qui ont une activité différente de celle normalement observée pour ce tissu (qui est nulle, rappelons le). Les effets de volume partiel se manifestant plus particulièrement à l'interface cortex/sillons, la fusion s'opère entre la carte de LCS issue de l'IRM et la carte de MG obtenue en TEM en recherchant les zones d'accord sur ces deux cartes [BARRA00-F]. L'opérateur retenu est le minimum, le plus grand des opérateurs conjonctifs. La carte floue ainsi créée représente les zones de volume partiel LCS/MG en TEM et est notée EVP. Cette carte EVP permet alors de reconstruire une nouvelle image TEM, corrigée des effets de volume partiel à l'interface cortex/sillons. L'activité *I(v)* d'un voxel de cette image est donnée par :

$$I(v) = \frac{b_{MG}.\pi^{MG-EVP}(v) + b_{MB}.\pi_B^{MB}(v) + b_{LCS}.\pi^{LCS}(v)}{\pi^{MG-EVP}(v) + \pi_B^{MB}(v) + \pi^{LCS}(v)},$$

où la distribution de possibilité π^{MG-EVP} est obtenue en effectuant la différence voxel à voxel entre les cartes floues MG obtenues en TEM et EVP, et où les pondérations b_T sont les activités caractéristiques des tissus dans l'image TEM issues de l'algorithme de classification. Les activités de l'image reconstruite expriment ainsi la même information que celles de l'image TEM originale. La Figure 1-5 présente une coupe d'une image TEM d'origine, les cartes EVP et MG corrigées et l'image TEM reconstruite.

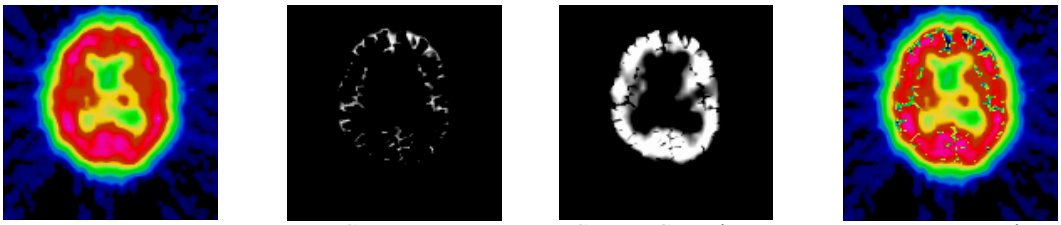


Image TEM | Carte EVP | Carte MG corrigée | Image TEM corrigée

Figure 1-5 : correction par fusion du volume partiel en TEM

3.2. Etude de l'épilepsie

L'épilepsie est la conséquence d'une modification brutale de l'électrogénèse du cerveau, et touche entre 80 000 et 400 000 personnes en France. Les principales conséquences sont une altération des mouvements, des sensations, des fonctions cognitives ou psychiques ou de la conscience. Il ne suffit cependant pas d'une seule crise pour devenir épileptique. C'est leur répétition au cours du temps qui caractérise la pathologie.

Dans le cadre d'une collaboration avec le Pr. Luc Cinotti et Nicolas Boussion (CERMEP, Lyon), nous avons souhaité montrer que la fusion d'images multimodales devait permettre de préciser le diagnostic de cette pathologie. Nous disposions pour cela d'un quadruplet d'images (IRM, TEM_c, TEM_i, TEP) pour 7 patients. L'image par résonance magnétique fournit ici la donnée morphologique, les images tomoscintigraphiques (acquises avec l'ECD comme traceur de perfusion) donnent accès au débit sanguin régional hors crise (TEM_i) et en crise (TEM_c), et l'image par tomographie par émission de positons (par injection de FDG) renseigne sur le métabolisme du glucose dans le cerveau.

3.2.1. Premiers résultats

Le type de fusion développé dans [BARRA01-D] a permis d'aboutir entre autres à une carte étiquetée, représentant les zones de forte possibilité de présence du foyer épileptogène primaire, et l'extension d'hypométabolisme du glucose associée, et par ailleurs démontrée dans la littérature (Figure 1-6).

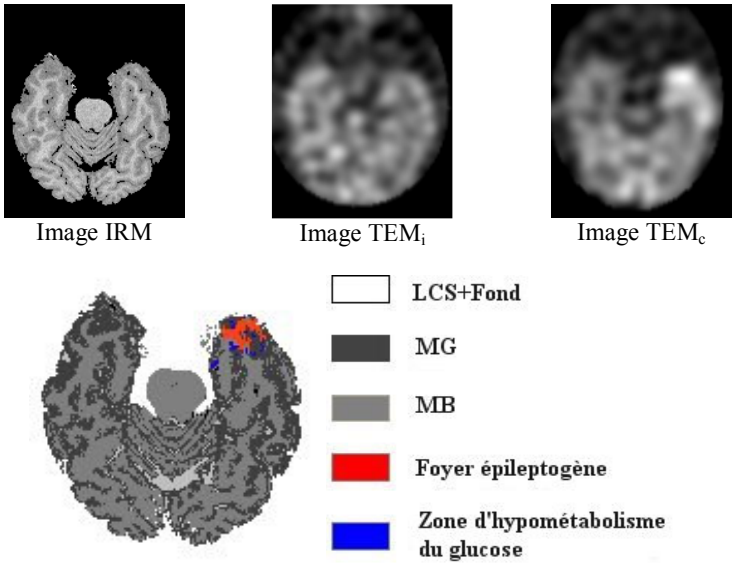


Figure 1-6 : image étiquetée pour l'épilepsie

3.2.2. Nouveaux résultats

L'intérêt de la méthode pour ce type particulier d'études m'a amené à poursuivre la collaboration avec Nicolas Boussion, maintenant en poste dans le laboratoire de recherche en imagerie et orthopédie de l'École de technologie supérieure de l'Université du Québec. Les travaux ont porté sur 12 patients qui, en plus des données précédentes, disposaient d'une TEP au Flumazenil permettant d'apprécier la densité d'inhibiteurs de la neurotransmission. Les résultats du processus de fusion sur ces données [BOUSSION03] se sont révélés en accord avec une référence clinique, représentée soit par une étude parallèle suite à l'implantation d'électrodes, soit par une étude post-chirurgicale suite à l'exérèse du foyer épileptogène. La Figure 1-7 présente, à titre d'exemple, quelques résultats du processus de fusion sur un des patients, par l'intermédiaire des images de synthèse IRM/TEM$_C$ (A), IRM/TEM$_i$ (B), IRM/PET$_{FDG}$ (A) et des zones anormales révélées par l'ensemble des images précédentes (D).

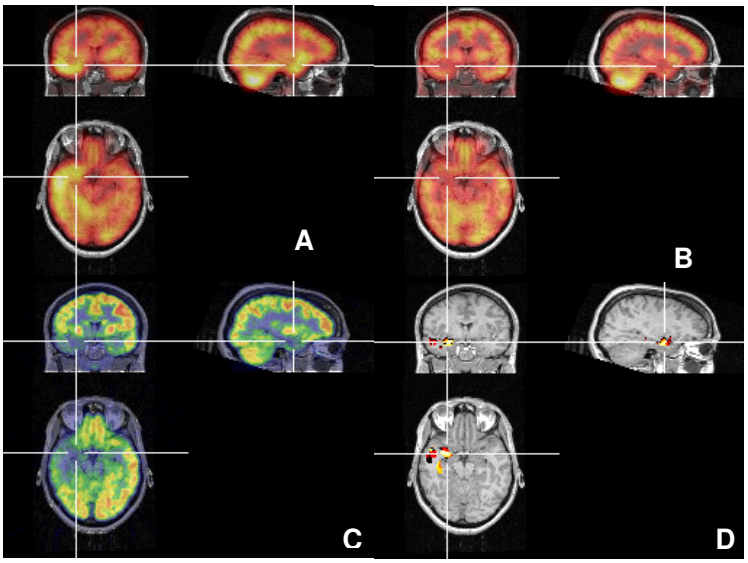


Figure 1-7 : image étiquetée pour l'épilepsie

Conclusion

Introduit dès 1997 au sein de l'Equipe de Recherche en Imagerie Médicale, et concrétisé par les premières applications en 2000, le concept de fusion tel qu'envisagé ici nous a semblé porteur de nombreuses applications potentielles, que ce soit pour l'étude multimodale d'une pathologie donnée ou pour l'agrégation d'images et d'informations expertes. Dans les chapitres suivants, les perspectives entrevues lors de mon travail de thèse vont être développées, à travers les travaux de trois doctorants que j'encadre actuellement, ou qui ont récemment soutenu.

Chapitre 2 - Segmentation de structures cérébrales

Cerveau: appareil avec lequel nous pensons que nous pensons.
Ambrose Bierce (Le dictionnaire du diable).

Introduction
La localisation fine des structures cérébrales profondes est un enjeu clinique essentiel, puisque ces dernières interviennent dans de nombreux mécanismes de fonctionnement du cerveau, et par là même participent à l'explication de l'étiologie de certaines pathologies. Le problème de segmentation sous-jacent reste cependant relativement difficile, en raison entre autres des faibles contrastes proposés dans les images, de la spécificité et de la complexité de chaque sujet étudié et de la taille caractéristique des régions d'intérêt. Ce chapitre décrit l'utilisation des concepts de fusion précédemment introduits comme outils de recherche d'une méthode de segmentation générique et automatique de structures cérébrales. Après avoir rappelé les intérêts d'une telle segmentation, et introduit quelques outils déjà existants, la modélisation générique du processus de segmentation de ces structures telle que nous l'envisageons est décrite. L'accent est en particulier mis sur la modélisation de l'ensemble des données disponibles, images et issues de conversations avec un expert, dont la coopération et l'agrégation aboutit à un schéma de fusion et de segmentation. De ce schéma sont ensuite dérivés des scénarii spécifiques à chaque structure d'intérêt. Ces scenarii sont alors enfin utilisés d'une part pour valider la méthode, et d'autre part comme supports à l'étude de pathologies en relation avec des services cliniques.

1. Segmentation de structures cérébrales

1.1. Intérêts

Une segmentation précise des structures cérébrales profondes est une problématique fondamentale rencontrée dans de nombreuses applications, allant de l'évaluation d'anomalies cérébrales structurelles [SHENTON92] à l'étude d'entités pathogènes (*e.g.* carcinomes [FILIPEK91], sclérose en plaques [KAMBER95], lésions de l'hippocampe [PRUESSNER00]), en passant par le mapping d'activations fonctionnelles sur l'anatomie cérébrale [MAZZIOTTA91], l'étude de pathologies impliquant des structures profondes (schizophrénie [IOSIFESCU97], hydrocéphalie [BRANDT94], maladie de Parkinson [BARRA01-B]) ou encore la neurochirurgie assistée par ordinateur [KIKINIS96].

1.2. Etat de l'art

De nombreuses méthodes semi- ou complètement automatiques ont été décrites dans la littérature pour réaliser cette segmentation, la plupart identifiant les structures d'intérêt en déformant les images natives, des références de forme ou des atlas à l'aide de déformations élastiques [HARTMANN99] [IOSIFESCU97] [KELEMEN99] [SZEKELY96] [THOMPSON97] ou de modèles déformables [DUTA98] [GHANEI98]. Ce type d'approche est efficace et fiable lorsqu'il s'agit d'effectuer de petites déformations locales, mais peut échouer dans le cas par exemple de grandes déformations [BAJCSY89], de faibles contrastes ou de trop petites structures à segmenter. Ces techniques sont également les plus souvent coûteuses à mettre en œuvre. D'autres approches fondées sur des algorithmes à base de règles [DELLEPIANE92] [LI95], de méthodes de traitement d'images (morphologie mathématique [DOKLADAL03], réseaux de neurones, [MAGNOTTA99], analyse d'histogramme [WORTH98], champs de Markov [HELD97], croissance de régions contrôlée [SAEED02], algorithmes de classification [WANG02]) ou coopérations de méthodes [JANG97] [SONKA96]) sont également proposées. Mais la plupart des techniques effectives en segmentation de structures se fondent finalement sur un tracé manuel du contour des régions d'intérêt, guidé [HARTMANN99] [NOWINSKI01] [YOTSUTSUJI03] ou non [GUNNING98] [HAMPEL02] [HEINZ92]

[TSATSANIS03] par un atlas, et sont donc sujettes aux variabilités intra- et inter-observateur, à un manque de reproductibilité certain et à une réalisation fastidieuse pour de larges échantillons à traiter.

2. Méthode proposée

J'ai initialement proposé dans [BARRA01-A] une méthode entièrement automatique de segmentation de structures cérébrales profondes, fondée sur un principe de fusion d'informations. Les données sont fournies à la fois par une image IRM et des connaissances expertes décrivant des relations topologiques, morphologiques et constitutives des structures d'intérêt.
La méthode repose sur les trois phases de fusion d'informations rappelées dans le chapitre 1. Dans un premier temps, les données disponibles sont modélisées dans le cadre possibiliste, pour ensuite être agrégées en tenant compte de leur redondance, de leur complémentarité et de leur conflit, ce qui permet finalement dans un dernier temps de prendre une décision quant à la localisation des structures à segmenter. L'ensemble de ce processus a été repris et approfondi lors du travail de thèse d'Emmanuelle Frenoux, [FRENOUX03-B] dont j'assurai le co-encadrement avec Jean-Yves Boire (thèse CIFRE SEGAMI, 2000-2003).

2.1. Préambule : segmentation d'amers

La modélisation abordée dans le paragraphe suivant suppose connue certaines structures remarquables appelés dans la suite amers. Il s'agit plus particulièrement ici du système ventriculaire (ventricules latéraux, V3 et V4) et du plan inter hémisphérique. Une méthode de définition de ces amers est donc tout d'abord présentée.

2.1.1. Segmentation du système ventriculaire

Sous l'hypothèse que le système ventriculaire est la plus grande composante connexe 3D du cerveau contenant du liquide cérébrospinal, son extraction suit le schéma suivant [BARRA02-B] : une segmentation du LCS est tout d'abord effectuée en utilisant l'algorithme développé dans [BARRA00-B], et la carte d'appartenance résultante est ensuite traitée en effectuant successivement (Figure 1-8) :

- une recherche d'un voxel *v* inclus dans le système ventriculaire : selon nos observations, la coupe maximisant le rapport volumique cerveau/fond contient toujours une trace des ventricules latéraux, et un tel voxel est choisi proche du centre de cette coupe avec un fort degré d'appartenance au LCS ;
- une recherche de la plus grande composante connexe 3D contenant *v* dans une 0.5-coupe de la carte d'appartenance au liquide ;
- une érosion morphologique 3D de l'image résultante par un élément structurant sphérique E de rayon 1.5mm, de manière à ne retenir que les points les plus surs de la composante connexe. Le rayon de E a été choisi pour creuser profondément les structures binaires et ainsi supprimer de petites composantes ayant un fort degré d'appartenance au LCS (*e.g.* sillons) ;
- une recherche de la plus grande composante connexe de l'image érodée ;
- une dilatation conditionnelle 3D par E ;
- une intersection de l'image résultante avec la 0.5-coupe de la carte d'appartenance au liquide, pour éviter une dilatation conditionnelle trop importante.

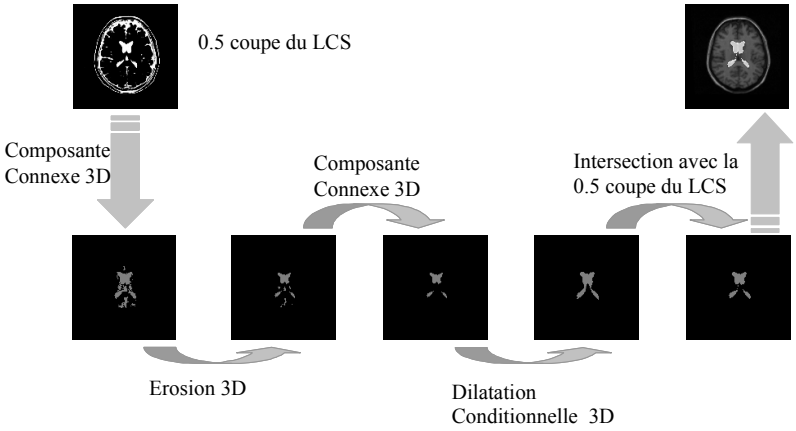


Figure 1-8 : segmentation du système ventriculaire

2.1.2. Définition du plan inter-hémisphérique

Le plan inter-hémisphérique étant par définition le plan de symétrie du cerveau, ce dernier est détecté tout d'abord en calculant les axes principaux d'inertie du système ventriculaire, puis en précisant la position des deux vecteurs de base du plan inter-hémisphérique ainsi déterminé par une maximisation de la corrélation gauche/droite du cerveau.

2.2. Modélisation des données

2.2.1. Données images

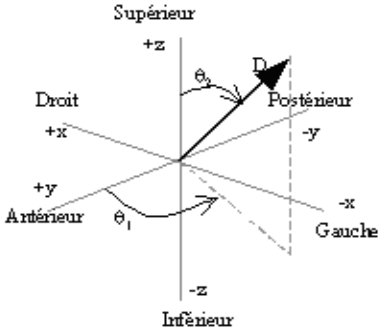
L'information numérique est directement extraite des images IRM par l'intermédiaire de l'algorithme de classification sur des vecteurs de coefficients d'ondelettes [BARRA00-B]. Le résultat est un ensemble de distributions de possibilité donnant pour chaque voxel v son degré d'appartenance $\pi_T(v)$ aux classes $T \in$ {fond, LCS, MB, MG}.

2.2.2. Informations topologiques

Les informations topologiques sont issues de la connaissance d'un expert et concernent la caractérisation en distance et direction des structures d'intérêt par rapport à des points de repère cérébraux.

- Représentation de l'information de direction

Bloch *et al* [BLOCH03] proposent une revue des méthodes de représentation des directions dans l'espace, ainsi qu'une comparaison des ces différentes méthodes. Nous nous intéressons ici à une description des informations topologiques à l'aide de propositions floues, et définissons pour cela six variables de position (Inférieur -I, supérieur -S, Antérieur -A, postérieur -P, Droite -D et gauche -G) dont l'ensemble des combinaisons d'au plus trois variables non antagonistes permet de décrire un 26-voisinage dans l'espace. Une direction D, donnée par une des combinaisons, est codée en coordonnées sphériques à l'aide des angles θ_1, formé par l'axe OY et la projection de D sur le plan OXY, et θ_2, l'angle formé par l'axe OZ et D (Figure 1-9).



A = (0,π/2), P = (π,π/2), G = (π/2,π/2), D = (-π/2,π/2),
S = (0,0), I = (0,π), AG = (π/4,π/2), AD = (-π/4,π/2),
PG = (3π/4,π/2), PD = (-3π/4,π/2), AS = (0,π/4),
AI = (0,3π/4), PS = (π,π/4), PI = (π,3π/4),
IG = (π/2,3π/4), ID = (-π/2,3π/4), SG = (π/2,π/4),
SD = (-π/2,π/4), AGS = (π/4,π/4), AGI = (π/4,3π/4),
PGS = (3π/4,π/4), PGI = (3π/4,3π/4),
ADS = (-π/4,π/4), ADI = (-π/4,3π/4),
PDS = (-3π/4,π/4), PDI = (-3π/4,3π/4).

Figure 1-9 : codage de la direction en 26 voisinage

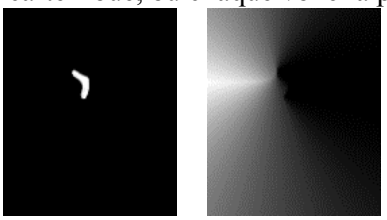
Une distribution de possibilité π_D est alors calculée à partir de cette description, en associant à chaque voxel v de l'image un degré d'appartenance à l'ensemble « dans la direction D d'une structure de référence S», par :

$$\pi_D(v) = \begin{cases} 0 & \text{si } v \in S \\ Max\left[0{,}1 - \dfrac{2\beta_D(v)}{\pi}\right] & \text{sinon} \end{cases}$$

où par exemple

$$\beta_D(v) = \begin{cases} 0 & \text{si } v \in S \\ \underset{o \in S_1}{Min}\left[Arccos\left(\dfrac{\overrightarrow{ov}.D}{\|\overrightarrow{ov}\|.\|D\|}\right)\right] & \text{sinon} \end{cases}$$

Ici encore, π_D est représentée par une carte floue, où chaque voxel a pour valeur $\pi_D(v)$ (Figure 1-10).



Structure S « A gauche de S »

Figure 1-10 : exemple de carte floue de direction

- Représentation de l'information de distance

Il s'agit ici de représenter par une distribution de possibilité π_M une proposition du type "S_2 est à la distance $F(d)$ de S_1", où F est un modificateur linguistique appliqué à la distance d. Cette modélisation passe tout d'abord par un calcul dans l'espace image d'une distance des voxels à S_1 (par exemple en utilisant une approximation du chamfrein), puis par une transformation de la carte de distance résultante en une carte floue portant la distribution π_M. Pour ce faire, chaque distance δ lue sur cette carte se voit affecter un degré d'appartenance au sous-ensemble flou « à $F(d)$ », à partir d'une représentation de la fonction d'appartenance de cet ensemble.

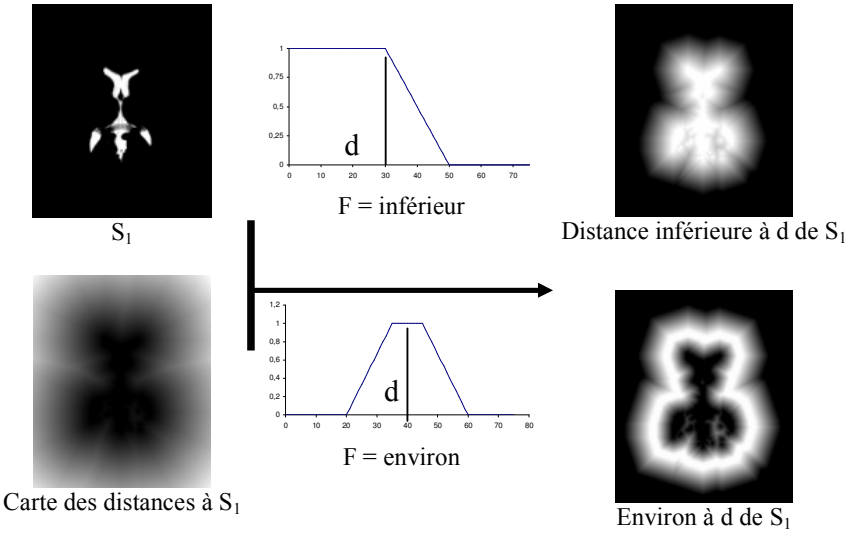


Figure 1-11 : exemple de carte floue de distance

2.2.3. Informations morphologiques

A la méthode originale publiée en 2001 [BARRA01-A], nous avons ajouté lors de la thèse d'Emmanuelle Frenoux [FRENOUX03-B] l'intégration d'une information morphologique sous la forme d'une représentation floue de la structure à segmenter. Une étude bibliographique, par ailleurs réalisée de manière fort complète dans [GERAUD98] et [GIBAUD97], montre que l'utilisation d'atlas est courante en segmentation de structures, et que la variété proposée est grande [SCHALTENB77] [TALAIRACH88]. En particulier, les atlas de type probabiliste affectent à chaque voxel une probabilité d'appartenance à un certain nombre de structures [MAZZIOTTA95] [NIEMANN95], et dans certains cas la variabilité de ces dernières peut être prise en compte en utilisant d'une part un squelette moyen, et d'autre part un espace de forme défini à partir d'un ensemble d'apprentissage et de surfaces paramétriques [STYNER03].

Nous avons naturellement cherché à produire pour chaque structure d'intérêt *S* une distribution de possibilité décrivant l'appartenance de chaque voxel à *S*. Quarante huit acquisitions IRM de 16 sujets sains et de 32 pathologiques (pathologies ne provoquant pas de variations significatives dans la morphométrie des structures d'intérêt) des deux sexes et d'âge variable ont pour cela été utilisées. Les structures d'intérêt ont été manuellement segmentées sur les 48 images avant recalage des IRM sur une IRM de référence (logiciel AIR [WOODS92]). Pour chaque structure *S*, une probabilité a alors été calculée en comptant la fréquence d'apparition de chaque voxel dans *S*, l'ensemble formant alors une carte floue d'appartenance à *S* (Figure 1-12).

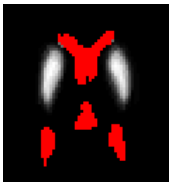


Figure 1-12 : carte de forme floue des putamens (en gris foncé, le système ventriculaire)

Cette carte est positionnée de façon automatique au cours du processus de segmentation. Les structures d'intérêt étant très proches du système ventriculaire, la mise en correspondance de ce dernier entre une IRM et l'IRM de référence assure que les structures d'intérêt sont relativement bien positionnées, ce qui est suffisant pour un processus de fusion, d'autres informations venant corriger l'imprécision de positionnement. Ce recalage élastique est effectué en deux temps, d'abord en utilisant une transformée en cosinus discrète [TOURAILLE00] pour mettre en correspondance les deux systèmes ventriculaires, puis en répercutant les paramètres de déformation sur les références de formes floues.

2.3. Fusion et décision

L'ensemble des cartes d'appartenance est ensuite agrégé en utilisant deux types d'opérateurs : un opérateur conjonctif F_c, généralisant l'intersection ensembliste, lorsque les informations sont redondantes (par exemple une carte de distance et une carte de tissu), et un opérateur disjonctif F_d, généralisant l'union d'ensembles, lorsque les cartes d'appartenance sont complémentaires (par exemple deux cartes de directions symétriques par rapport au plan inter hémisphérique). Pour un voxel v donné, de degrés d'appartenance $(v_i)_{1 \leq i \leq n}$ aux différentes données, ces opérateurs s'écrivent par exemple pour $n = 2$:

$$F_c(v_1,v_2) = v_1.v_2$$
$$F_d(v_1,v_2) = v_1 + v_2 - v_1.v_2.$$

Pour chaque voxel v, le résultat de l'étape d'agrégation est un degré d'appartenance $v_f = F(v_i)_{1 \leq i \leq n}$ synthétisant l'ensemble des informations disponibles, et gérant par l'intermédiaire de F_c et F_d leur redondance et leur complémentarité. Pour le cas qui nous intéresse ici, v_f est interprété comme le degré d'appartenance de v à la structure d'intérêt considérée, et la carte des v_f est une carte d'appartenance à cette structure. Une décision est alors prise sur cette carte en effectuant une 0.5 coupe du sous-ensemble flou sous-jacent, l'image résultante représentant la structure segmentée par le processus de fusion. La Figure 1-13 résume l'ensemble de ces étapes sur la segmentation des putamens.

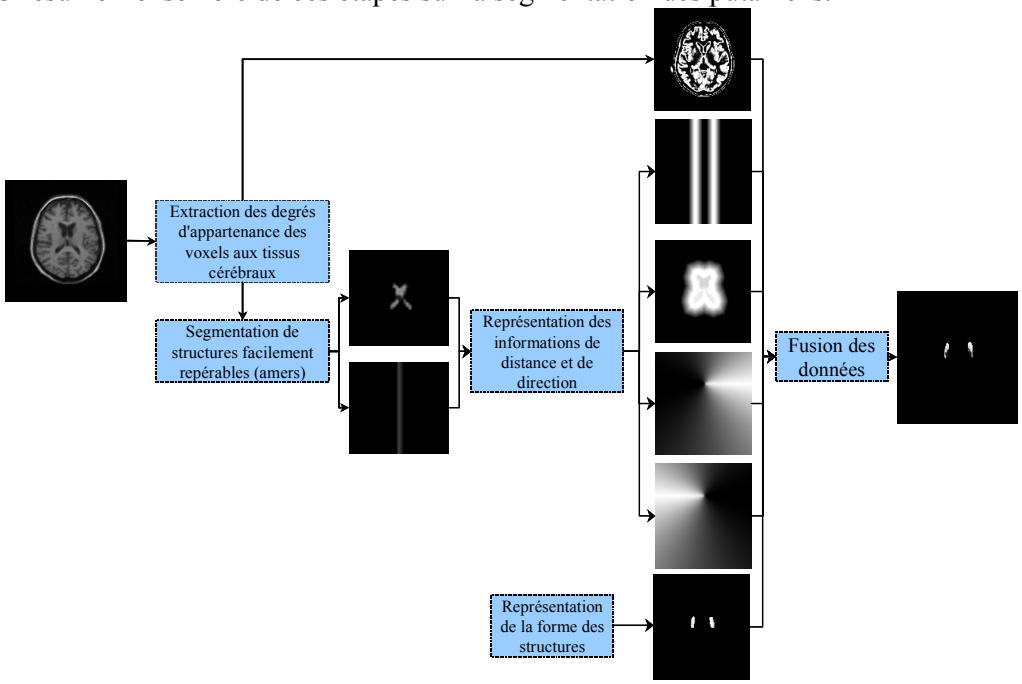


Figure 1-13 : processus de segmentation de structures

2.4. Déclinaison et écriture de scenarii

Ce schéma de segmentation générique a été appliqué lors de la thèse d'Emmanuelle Frenoux à l'extraction de structures d'intérêt dans le cas de diverses applications cliniques (*cf.* 3.2 et chapitre suivant). L'ensemble de ces travaux a été réalisé en collaboration avec le Dr Marie-Odile Habert, clinicien du service de médecine nucléaire de la Pitié-Salpêtrière à Paris, qui a joué le rôle d'experte. Plusieurs scenarii, variant par la nature et la quantité d'informations disponibles, ont été envisagés. Tous suivent néanmoins le même schéma algorithmique :

1. segmenter les amers
2. extraire les cartes d'appartenance aux tissus
3. représenter les informations topologiques par rapport aux amers
4. recaler la référence de forme dans le référentiel de l'image

5. effectuer la fusion entre les différentes informations
6. calculer la structure binaire segmentée.

2.4.1. Segmentation de la tête des noyaux caudés

Selon l'expert, les têtes de noyaux caudés sont collées à la corne frontale des ventricules latéraux (distance inférieure à 2mm), sont de part et d'autre de cette dernière et sont composées de matière grise. La référence de forme est modélisée sous la forme de la distribution de possibilité présentée sur la Figure 1-14.

Figure 1-14 : référence de forme des têtes de noyaux caudés

2.4.2. Segmentation des putamens

Les informations fournies par M.O. Habert précisent que les putamens sont à environ 28mm des ventricules latéraux, à environ 72 mm du plan inter hémisphérique, en direction postérieure gauche et postérieure droite de la corne frontale des ventricules latéraux, et sont composés de matière grise. La référence de forme est modélisée sous la forme de la distribution de possibilité présentée sur la Figure 1-15

Figure 1-15 : référence de forme des putamens

2.4.3. Segmentation du thalamus

Le thalamus est collé au troisième ventricule, en direction postérieure droite et postérieure gauche par rapport à ce dernier, est composé de matière grise et a pour référence de forme la structure présentée sur la Figure 1-16.

Figure 1-16 : référence de forme du thalamus

2.4.4. Segmentation du cervelet

Le cervelet est à une distance supérieure à 100mm et inférieure à 160mm du plancher des ventricules latéraux, en direction postérieure et inférieure par rapport à ces derniers. Composé d'un mélange de matière blanche et de matière grise, la carte de tissu pour cette structure a été calculée en considérant que les voxels ayant un degré d'appartenance non nul à la MG et à la MB mais également un degré d'appartenance élevé pour le LCS sont mis à zéro au cours du processus de fusion, ce qui permet de retirer certains sillons assez étroits pénétrant dans le cervelet. La référence de forme quant à elle est donnée sur la Figure 1-17.

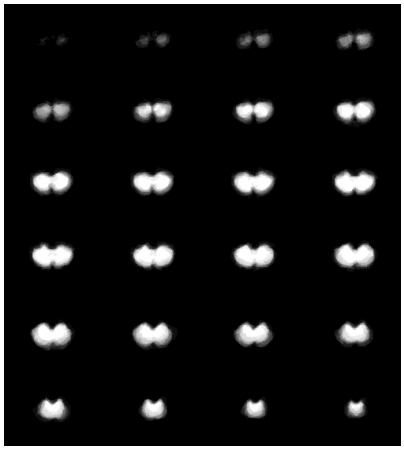


Figure 1-17 : référence de forme du cervelet

2.4.5. Sectorisation du cortex

Trois parties distinctes du cortex sont ici étudiées (frontal, occipital et temporal), toutes contenues dans les coupes contenant le système ventriculaire. Ces structures sont à une distance supérieure à 10mm du crâne, lui-même extrait à partir d'un masque du fond, et supérieure à 30 mm du système ventriculaire. Le cortex frontal est en direction frontale par rapport au système ventriculaire, le cortex temporal à droite et gauche de ce dernier et la partie temporale est en direction postérieure. Toutes trois sont composées de matière grise, et aucune référence de forme n'a été ici calculée, le cortex étant formé de replis de matière grise complexes.

3. Résultats et discussion

3.1. Validation

Les résultats obtenus pour la segmentation automatique ont été dans un premier temps validés visuellement par le Dr M.O. Habert, puis comparés à ceux obtenus à l'issue d'un tracé manuel par l'expert sur 48 images IRM. Trois indices numériques ont pour cela été définis [BARRA01-A] : le premier est un indice de similarité, calculé à partir de l'erreur relative d'estimation des volumes :

$$I_1 = 1 - \frac{|V_M - V_C|}{V_M}$$

où V_M est le volume obtenu manuellement, et V_C le volume calculé à partir du résultat de la segmentation automatique. Le second indice permet d'apprécier le recouvrement spatial des deux structures de référence S_M et segmentée S_C :

$$I_2 = \frac{Card(S_M \cap S_C)}{Card(S_M)}$$

Enfin, le dernier indice évalue la distance moyenne entre S_M, et S_C par :

$$I_3 = \frac{\sum_{P_C \in S_C} \underset{P_T \in S_T}{Min} \| P_C P_M \|}{Card(S_C)}$$

Ces trois indices ont été calculés pour les têtes des noyaux caudés, les putamens et le cervelet [FRENOUX02-A], mais pas pour les zones corticales segmentées, en raison de la difficulté de réaliser une segmentation manuelle de ces structures. La validation du processus de segmentation du thalamus est en cours de traitement. A titre d'illustration, les valeurs moyennes des indices I_1 et I_2 latéralisés sont présentées dans le Tableau 1-2.

	$\overline{I_1} \pm \sigma$	$\overline{I_2} \pm \sigma$
Têtes de noyaux caudés gauche / droite	0,92±0,02 / 0,90±0,04	0,85±0,06 / 0,84±0,05
Putamens gauche / droit	0,94±0,03 / 0,93±0,03	0,88±0,03 / 0,88±0,04
Cervelet	0,94 ± 0,04	0,83 ± 0,03

Tableau 1-2 : valeurs moyennes des indices de validation de la segmentation

La distance moyenne I_3 calculée pour les têtes de noyaux caudés et les putamens est inférieure à 2mm pour 90% des structures et le plus mauvais résultat obtenu est une distance de 3mm. Cette mesure nous a par contre semblé peu pertinente pour la validation de la segmentation du cervelet, ce dernier présentant de nombreux sillons intégrés dans le contour manuel (faute d'un outil permettant de les retirer correctement et d'une tâche fastidieuse à effectuer), mais pas dans la structure segmentée automatiquement. Nous avons donc proposé [FRENOUX04] un quatrième indice I_4, évaluant les différences d'étiquetage entre S_C et S_M :

$$I_4 = \frac{Card(S_M \cup S_C) - Card(S_M \cap S_C)}{Card(S_M)}$$

Le pourcentage d'erreurs d'étiquetage dans le cervelet s'avère assez élevé (25,6%), du fait du retrait des sillons par l'algorithme. Afin de confirmer cette hypothèse, nous avons calculé le pourcentage de sillons contenus dans la surface obtenue manuellement. Pour ce faire, nous avons réalisé l'intersection entre la carte de liquide cérébrospinal de l'image et la segmentation manuelle (Figure 1-18) : 21% des points intégrés au cervelet appartiennent en réalité au LCS, ce qui signifie que les voxels "mal classés" au cours du tracé manuel (en blanc sur la figure) représentent environ 82% des erreurs d'étiquetage relevées, et s'avèrent donc être des améliorations par rapport à la méthode manuelle.

Figure 1-18 : correction de la segmentation du cervelet

Reproductible et entièrement automatique, le processus proposé ici se fonde sur le cumul d'informations hétérogènes, redondantes et complémentaires pour proposer une segmentation robuste et conforme aux résultats fournis par l'expert. La méthode est de plus flexible, autorisant la segmentation d'un grand nombre de structures cérébrales, pourvu que l'on dispose d'informations topologiques et/ou

morphologiques et que l'acquisition de référence ait une résolution et un contraste permettant d'identifier la cible à segmenter.
Ce processus de segmentation, par ailleurs publié sur la segmentation de diverses structures [BARRA01-A] [BARRA01-B] [FRENOUX02-B] [FRENOUX03-A] [FRENOUX04], a été porté sur les consoles Mirage de la société SEGAMI et fait actuellement l'objet d'une demande de dépôt de brevet (Annexe B).

3.2. Applications

Plusieurs applications cliniques ont naturellement découlé de ce processus de segmentation. Les applications à visée purement morphologiques sont présentées dans la suite. Dans le chapitre 3, les applications utilisant ce type de segmentation pour des problèmes de quantification fonctionnelle seront plus spécifiquement développées.

3.2.1. Maladie de Parkinson

La maladie de Parkinson idiopathique (MPI) est une affection neurodégénérative du système nerveux central touchant environ 100.000 personnes en France, deuxième par ordre de fréquence après la démence de type Alzheimer. Cette affection débute entre 55 et 60 ans et entraîne un ralentissement gestuel, une rigidité et un tremblement de repos. Les signes cliniques sont en rapport avec une diminution des concentrations en dopamine dans le striatum, liée à la perte des neurones dopaminergiques de la substance noire. Si un traitement médicamenteux (L.dopa) donne des résultats satisfaisants pendant les premières années, des complications apparaissent au cours de l'évolution de la maladie, qui sont à l'origine d'une perte progressive de l'autonomie des patients. Un effort considérable a été fait pour comprendre l'organisation fonctionnelle des noyaux gris centraux à l'état normal et les modifications de leur fonctionnement après dénervation dopaminergique. Dans un modèle développé au début des années 90 [VILA96], le noyau sous-thalamique (NST) joue un rôle central, de par son hyperactivité dans le modèle du primate rendu parkinsonien et par le rôle bénéfique d'une stimulation ou d'une lésion chirurgicale sur la symptomatologie parkinsonienne. Quel que soit l'acte chirurgical envisagé, il est donc essentiel de pouvoir repérer efficacement cette structure en pré ou per opératoire, et j'ai proposé en 2001 [BARRA01-B] d'effectuer une segmentation du NST utilisant le processus de segmentation développé dans ce chapitre, dans le cadre de la stimulation par électrodes de ces noyaux.

Le Pr. Jean-Jacques Lemaire, neurochirurgien au CHU de Clermont-Ferrand, est l'expert dans cette étude, qui intéresse 7 patients pour lesquels trois séquences d'images IRM orthogonales de 17 coupes jointives de 256×256×3mm^3 ont été acquises en conditions stéréotaxiques (cadre Leksell Modèle G et système de repositionnement à 4 fixations corticales, Elekta Instruments®). L'ensemble des informations nécessaires au déroulement de l'algorithme a été fourni par l'expert, selon le scénario suivant (Figure 1-19) :

- détermination de la position du noyau rouge (NR) à partir du troisième ventricule (V3), par des informations de niveaux de gris, de distance et de direction ;
- détermination de la position du noyau sous-thalamique à partir du noyau rouge, à l'aide d'informations de niveaux de gris, de distance et de direction.

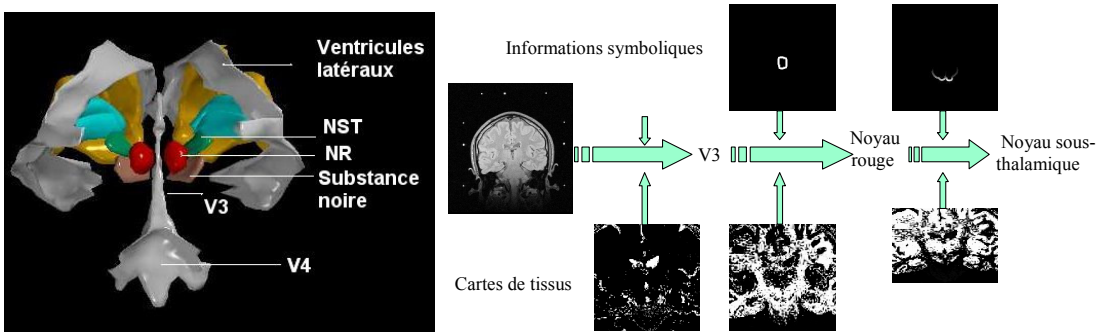


Figure 1-19 : scénario de segmentation du noyau sous-thalamique

Un exemple de résultat est présenté sur la Figure 1-20. Une étude plus approfondie de l'anatomie fine autour du NST montre en fait que la structure segmentée par l'algorithme est un agrégat du NST et du locus niger, qui forment sur l'image IRM une seule structure complexe de matière grise.

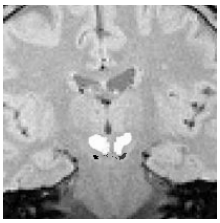


Figure 1-20 : segmentation du NST (en blanc : noyau rouge ; en noir : NST)

La validation de cette segmentation passe par une confrontation des résultats calculés avec des données cliniques, disponibles sous trois formes : les coordonnées radiographiques, mesurées par le neurochirurgien sur un cliché de contrôle en fin de procédure chirurgicale. Ces coordonnées sont celles du plot de l'électrode qui stimule effectivement le NST, et seront considérées comme la référence ; les coordonnées dans le repère de Talairach, relevées sur l'IRM stéréotaxique; et les coordonnées de Benabid, relevées sur l'IRM stéréotaxique. Une étude préliminaire [BARRA01-B] a montré des résultats encourageants, et a souligné les points d'amélioration possibles de la méthode. Une validation complète et une reprise de l'algorithme sur de nouveaux jeux de données sont actuellement en cours dans le cadre d'une thèse en contrat MENRT à l'ERIM (Alice Villéger).

3.2.2. Sectorisation du cortex pour l'étude de la démence de type Alzheimer

Le processus de sectorisation automatique proposé ici a été mis en place en collaboration avec le Pr. Michèle Allard (hôpital Pellegrin, Bordeaux) afin d'automatiser le processus de quantification mis en place dans son service pour l'étude de la démence de type Alzheimer (DTA). Dans ce cadre, deux acquisitions sont réalisées, l'une morphologique (IRM 3D SPGR) donnant une référence anatomique, et l'autre fonctionnelle (TEM, tomoscintigraphie avec un traceur de perfusion) permettant d'apprécier l'atteinte de la pathologie. Le but est ici de proposer une segmentation rapide et reproductible du cortex, indépendante de l'opérateur, et permettant de donner des zones anatomiques de référence dans lesquelles quantifier en TEM l'évolution de la pathologie. Onze secteurs corticaux par hémisphère sont plus précisément concernés :

- le cortex orbito-frontal (droit et gauche)
- le cortex dorso-latéral frontal droit et gauche ;
- le gyrus cingulaire antérieur droit et gauche ;
- le gyrus cingulaire postérieur droit et gauche ;
- le cortex primaire droit (CPD) et gauche (CPG) ;
- le cortex pariétal supérieur (droit (CPSD) et gauche (CPSG) ;
- le gyrus temporal supérieur droit et gauche ;
- le gyrus temporal moyen droit et gauche ;
- le gyrus temporal inférieur droit et gauche ;
- l'hippocampe et le gyrus para-hippocampe droit et gauche ;
- le cortex occipital droit et gauche ;

A la partition du cortex s'ajoutent 6 structures cérébrales supplémentaires :

- le thalamus droit et gauche ;
- le striatum droit et gauche ;
- le cervelet droit et gauche .

Au total, 28 structures sont donc à segmenter. Parmi ces dernières, certaines serviront de régions non spécifiques, car non atteints par la DTA et non fixés par le traceur (*e.g.* le cervelet), tandis que d'autres

seront atteints en fonction du degré d'avancement de la pathologie (par exemple le gyrus cingulaire antérieur, touché seulement lorsque la DTA est très avancée ou l'hippocampe, première zone touchée), et que certaines enfin ne seront pas touchées par la DTA (*e.g.* le striatum), mais permettront de diagnostiquer une pathologie différente telle que les syndromes parkinsoniens, qui peut évoluer en même temps que la DTA.
La méthode de segmentation utilise ici seulement les cartes de tissus et des modèles flous des secteurs corticaux, réalisés à partir de segmentations manuelles effectuées par le Dr. Adriana Petrescu (Figure 1-21).

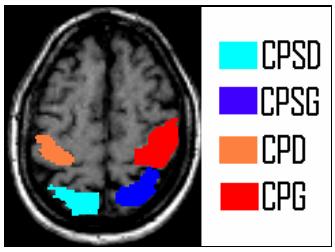


Figure 1-21 : segmentation manuelle de 4 secteurs corticaux sur une coupe d'IRM 3D

Le processus de sectorisation automatique du cortex a été appliqué à sept images IRM (3DSPGR, 256×256×128 coupes, pour une taille de voxel de 1mm et une épaisseur de coupes de 1,7mm), à l'issue duquel 28 secteurs de MG et autant de LCS sont obtenus (Figure 1-22).

Figure 1-22 : exemple de résultat de l'extraction de la MG contenue dans le CPSD

Le résultat a été comparé à la sectorisation manuelle. Toutefois, les secteurs extraits manuellement sont extrêmement larges et englobent non seulement la MG mais aussi le LCS et la MB alors que le but de la sectorisation automatique est d'extraire la MG de ces secteurs. Deux nouveaux indices de comparaison ont donc été utilisés : l'intersection entre le secteur automatique et la partie de matière grise inclue dans la région correspondante dessinée à la main (obtenue par une intersection entre le tracé manuel et la carte d'appartenance à la MG) et le pourcentage de points du secteur automatique se trouvant à l'extérieur de la région manuelle correspondante.
Ainsi, pour les secteurs extraits en MG, on obtient une intersection moyenne de 92.2% et un pourcentage moyen de points hors secteurs de 5.6%. Pour les secteurs extraits en LCS, on obtient une intersection moyenne de 83.3% avec un pourcentage moyen de points hors secteurs de 8.4%. Le volume obtenu à l'issue de la sectorisation automatique est inférieur d'environ 30% à celui obtenu par l'intersection entre la région dessinée manuellement et la carte de MG, ce qui s'explique, entre autres, par la taille des régions manuelles, qui sont souvent sur-dimensionnées, et entraînent l'inclusion de point aberrants dans les régions d'intérêt.
La Figure 1-23 présente un exemple de résultat obtenu pour le cortex primaire et pour le striatum, avec en regard l'intersection entre la référence manuelle et la carte de MG, la référence manuelle et l'IRM de référence. Les points aberrants ont été entourés dans les deux cas.

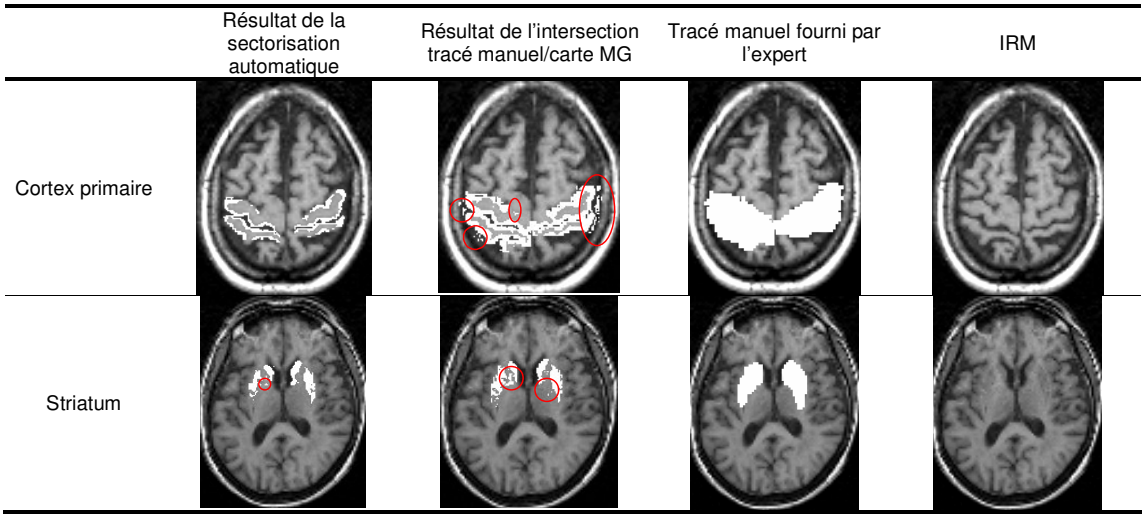


Figure 1-23 : comparaison entre la sectorisation automatique et la sectorisation manuelle

Conclusion

La méthode proposée ici est une technique de segmentation générique de structures cérébrales, dont les seules exigences sont un contraste suffisant sur l'imagerie de référence, et des connaissances expertes permettant de caractériser la position des structures d'intérêt les unes par rapport aux autres. Validée sur un certain nombre de structures impliquées dans diverses pathologies, cette méthode est actuellement à l'étude pour l'aide au positionnement d'électrodes de stimulation profonde dans le cadre de la maladie de Parkinson. Les perspectives d'utilisation de cette technique sont très nombreuses, et incluent la continuation des études en cours, la recherche d'autres structures cibles en relation avec les cliniciens et la construction des scenarii correspondants.

Il est évidemment également envisageable d'étendre le champ d'application de cette technique à d'autres études, par exemple concernant la quantification en imagerie multimodalité pour l'étude de pathologies neurodénégératives, ce qui fait l'objet du chapitre suivant.

Chapitre 3 - Quantification en imagerie cérébrale

La valeur d'une image se mesure à l'étendue de son auréole imaginaire.
Gaston Bachelard.

Introduction

La morphométrie des structures cérébrales n'est pas seule atteinte au cours du développement de certaines pathologies. Le fonctionnement cérébral est également souvent altéré, soit par la dégradation de certains tissus, soit par une altération des fonctions de certaines structures. Toute la difficulté repose alors sur une localisation précise de ces dégradations. Ce chapitre présente deux approches permettant de quantifier ces altérations dans le cadre clinique de pathologies données. La première exploite le processus de segmentation précédemment développé comme outil de repérage des informations pertinentes dans des données fonctionnelles. Utilisant un ensemble d'images anatomiques et fonctionnelles recalées entre elles, elle permet la recherche des zones d'intérêt dans l'image anatomique pour une quantification précise des zones d'activité correspondantes dans les images fonctionnelles. La seconde revient sur le processus de fusion d'images anatomiques et fonctionnelles envisagé dans le premier chapitre, et considère le caractère discret des pavages de voxels dans les images pour proposer de nouveaux développements issus de la géométrie discrète. Validées sur des fantômes numériques, ces méthodes sont ensuite appliquées à des données réelles, en collaboration avec divers services cliniques en France et dans le cadre d'un contrat RNTS (Réseau National Technologies de la Santé) débuté en 2002.

1. Quantification à partir de la segmentation

Dans le cadre de la thèse d'Emmanuelle Frenoux, nous avons mis en place un processus de quantification fonctionnelle automatique à partir des scénarii de segmentation présentés dans le chapitre 2. Le but ici est d'assister le clinicien dans le cadre de pathologies neurodégénératives étudiées à l'aide d'images anatomiques et fonctionnelles. Pour les exemples présentés dans ce chapitre, une image IRM et une ou plusieurs tomoscintigraphies sont disponibles pour chaque sujet analysé, la première donnant accès à une localisation des régions et structures d'intérêt, et la (les) seconde(s) permettant d'évaluer un processus physiologique (perfusion sanguine, neurotransmission) dans ces régions.

1.1. Méthode proposée

Le processus de quantification développé dans cette partie utilise la méthode de segmentation automatique de structures cérébrales développée et validée dans le chapitre précédent. Contrairement aux techniques classiquement utilisées dans des études similaires, qui s'appuient le plus souvent sur un contourage manuel (guidé par un atlas par exemple) ou au mieux semi-automatique des structures, la méthode que nous proposons est robuste, totalement reproductible est insensible à la variabilité inter et intra opérateur. De plus, elle allège considérablement la tâche du clinicien par rapport par exemple à une segmentation manuelle et fastidieuse des régions d'intérêt sur des volumes de données importants.

1.1.1. Principe

Le processus de quantification se décompose en trois phases. La première étape consiste à recaler les images anatomiques et fonctionnelles dans un même référentiel géométrique. Pour ce faire, nous avons utilisé et validé avec le clinicien référent une technique de recalage rigide fondée sur l'algorithme de Woods [WOODS92]. La seconde étape se propose alors d'isoler dans l'image anatomique les structures d'intérêt préalablement nommées par l'expert en fonction de la pathologie donnée, par l'intermédiaire d'un scénario de segmentation similaire à ceux développés dans le chapitre précédent. Finalement, la dernière phase du processus consiste à quantifier les activités des régions d'intérêt, par l'intermédiaire d'indices numériques calculés dans les zones 3D issues de la segmentation anatomique et reportées sur les images fonctionnelles (Figure 1-24).

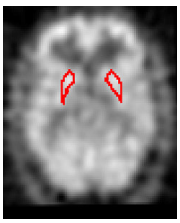


Figure 1-24 : report des structures segmentées sur une coupe de l'image fonctionnelle

1.1.2. Indices de quantification

Le calcul des indices numériques de quantification est effectué en rapportant les activités des régions d'intérêt à une valeur fonctionnelle non spécifique, issue d'une région où le traceur utilisé pour l'acquisition des images fonctionnelles n'est pas censé fixer. Différentes structures non spécifiques sont segmentées dans la suite en fonction du choix du clinicien et de celles relevées dans la littérature : le cervelet, le cerveau complet, le cortex complet ou divisé en trois zones (frontale, temporale, occipitale). La validité des indices présentés au clinicien dépendra en grande partie du choix de cette région non spécifique, et la méthode de segmentation proposée au chapitre précédent permet ici d'extraire de manière automatique et précise ces zones.

Dans le cadre des applications cliniques proposées ci-après, plusieurs indices ont été calculés dans les régions d'intérêt. Citons par exemple la fixation moyenne normalisée par la moyenne non spécifique (binding potential), ou par le maximum non spécifique, ou encore la fixation minimum (ou maximum) normalisée par la moyenne (ou le maximum) non spécifique. Le minimum et le maximum normalisés permettent en particulier une comparaison plus fiable inter patients ou pour une étude longitudinale menée sur le même sujet.

Chacun de ces indices est dans la suite calculé pour la structure gauche, la structure droite et l'ensemble des deux structures, pour prendre en compte les atteintes asymétriques de certaines pathologies (*e.g.* Parkinson), qui sera évaluée plus aisément avec des indices de fixation latéralisés.

Citons enfin deux indices qui peuvent être classés parmi les indices normalisés bien que n'utilisant pas explicitement la fixation observée dans la zone non spécifique : l'asymétrie de fixation signée, qui donne une évaluation de la latéralisation de l'atteinte en même temps que la fixation, et le rapport entre les fixations moyennes dans les noyaux caudés et dans les putamens, dont l'utilité a été démontrée pour l'étude de l'état d'avancement de la maladie de Parkinson ([LAULUMAA93]).

1.2. Applications

1.2.1. Sclérose latérale amyotrophique

La Sclérose Latérale Amyotrophique (SLA) est une pathologie impliquant la dégénérescence et la perte des neurones moteurs, et a une prévalence de 6 à 7 personnes sur 100 000 [ROWLAND01]. Mortelle et incurable, elle atteint peu à peu l'ensemble du système musculaire, et son diagnostic s'appuie essentiellement sur les symptômes cliniques et la négativité des examens complémentaires à la recherche d'une autre cause (tumeur,…). Les différentes études réalisées concluent à des défauts de perfusion dans le cortex moteur [WARAGAI97], et plus particulièrement dans la partie frontale du cortex [ABE97] [PORTET01]. Des observations réalisées simultanément sur des acquisitions IRM et fonctionnelles permettent de plus de mettre en évidence une atrophie frontale [ABE97] [ROWLAND01], et suggèrent également la possibilité d'une anomalie dans le circuit de neurotransmission dopaminergique.

Deux types de SLA sont reconnus, selon qu'elles sont raides (SLAR) ou non (SLANR). La raideur étant également une manifestation des syndromes parkinsoniens, le but de l'étude réalisée ici à terme est d'observer la fixation des traceurs spécifiques des récepteurs D2 de la dopamine dans le striatum (et plus particulièrement dans les têtes de noyaux caudés) afin de différencier, si possible, les sujets sains des sujets atteints de SLA, et de différencier également les deux types de SLA.

Sous l'expertise du Dr Marie-Odile Habert, 47 sujets appariés en âge ont été étudiés, comprenant 16 sujets sains (59.62 ± 10.24 ans), 13 sujets SLANR (55.15 ± 13.73 ans) et 18 sujets SLAR (51 ± 12.67

ans). Dans une étude préliminaire, deux acquisitions ont été réalisées pour chacun des sujets : une IRM pondérée en T_1 (128×128×128, 8mm^3) et une acquisition TEMP (traceur ^{99m}Tc-ECD de perfusion sanguine), recalée et redimensionnée sur l'IRM par l'expert. Les structures d'intérêt segmentées sont le striatum (dissocié ou non) d'une part, et des régions non spécifiques (cortex, cortex frontal, temporal et occipital, cervelet et cerveau complet) d'autre part.
Les volumes latéralisés et le volume total de la structure normalisé par le volume du cerveau ont été calculés pour les têtes des noyaux caudés (TNC), les putamens (PU), le striatum (ST) complet et pour les ventricules latéraux (VL). Des indices de fixation ont également été calculés pour les TNC, les PU et pour le striatum complet : la fixation moyenne, normalisée par le maximum et la moyenne de la zone non spécifique et non normalisée, l'asymétrie de fixation signée, le rapport de fixation NC/PU ont été choisis.

- Indices volumiques

Une analyse statistique par des tests de Kruskal-Wallis et de Mann et Whitney (avec correction de Bonferonni pour prendre en compte la différence d'effectifs) permet de faire ressortir une différence significative de volumes normalisés entre sains et SLA pour les VL ($p=0.01$) et le striatum ($p<0.00001$). Les autres variations de volume sont non significatives (Figure 1-25).

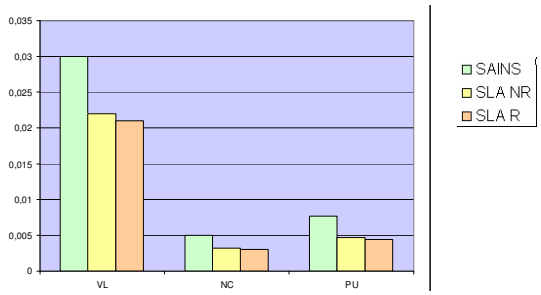


Figure 1-25 : volumes normalisés des structures d'intérêt pour la SLA

Bien que n'ayant pas trouvé dans la littérature d'études concernant les modifications de volume du striatum ou des VL par la SLA, les résultats ont été validés par l'expert [FRENOUX03-A].

- Indices quantitatifs

Un test de Kruskal-Wallis a permis de trouver des variations significatives de l'indice de fixation moyenne non normalisée sur la tête des noyaux caudés ($p=0.003$) et des indices normalisés d'asymétrie de fixation du putamen ($p=0.009$) et de fixation de la TNC par rapport au cervelet ($p=0.01$) (Figure 1-26)

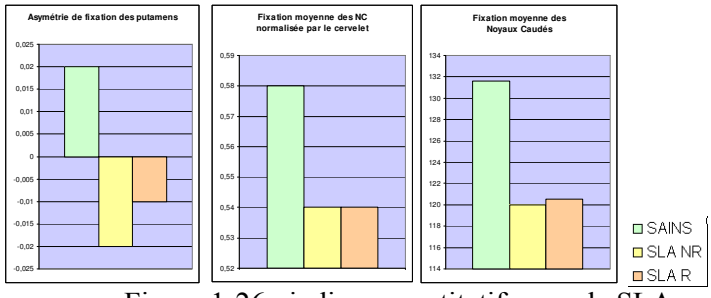


Figure 1-26 : indices quantitatifs pour la SLA

Globalement, l'atteinte du striatum est plus ou moins latéralisée suivant la forme prise par la SLA. Ainsi, si la fixation moyenne totale, gauche et droite pour les TNC est significativement différente entre sujets sains et SLAR, seules les fixations moyennes totale et droite sont significativement différentes entre sujets sains et SLANR. Les indices de fixation des TNC normalisés, quant à eux, ne font ressortir que la diminution de fixation dans la TNC droite, quels que soient les cas. De même, l'asymétrie de fixation des putamens est significativement différente dans tous les cas, mais la fixation moyenne dans le putamen

droit n'est pas significativement différente entre sujets sains et SLANR, tandis qu'elle l'est entre sains et SLAR. Toutefois, cette différence n'est pas observée sur la fixation normalisée des putamens, mais l'hypothèse peut malgré tout être faire que l'atteinte dans le putamen droit est plus importante chez les sujets SLAR.
Ainsi, sur l'étude proposée, les indices de fixation permettent de différencier sujets sains et sujets SLA (raides et non raides confondus), mais ne sont pas assez discriminants pour différencier les sujets pathologiques entre eux. D'ailleurs, aucune étude à ce jour n'a pour l'instant montré une atteinte significativement différente entre les deux états de la pathologie.

Nous avons d'ores et déjà débuté une seconde étude concernant la neurotransmission dopaminergique dans le cadre de la SLA. Pour cette dernière, des acquisitions cliniques sont en cours de réalisation, sur des volontaires sains, des sujets SLAR et SLANR. Deux traceurs de neurotransmission sont utilisés pour chaque sujet : l' ^{123}IBZM, marqueur des récepteurs D2 post-synaptiques, et le DaT-SCAN, un marqueur des récepteurs D2 pré-synaptiques. Cette étude vise à différencier SLAR et SLANR sous l'hypothèse que l'atteinte pré et post-synaptique n'est pas la même dans les deux cas. Elle permettra également de valider une nouvelle fois les indices calculés en comparant les résultats obtenus avec ceux de la littérature, dans laquelle par exemple Vogels *et al* [VOGELS00] concluent à une diminution de la concentration des récepteurs D2 de la dopamine causée selon eux à un excès de glutamate dans le striatum (dû à la SLA), qui inhibe la synthèse des récepteurs D2.

1.2.2. Maladie de Parkinson

Nous proposons ici une application des méthodes développées pendant la thèse d'Emmanuelle Frenoux au diagnostic différentiel entre maladie de Parkinson idiopathique (MPI) et différents syndromes parkinsoniens (atrophie multi système, paralysie supra nucléaire progressive), ces derniers réagissant peu au traitement à la L-Dopa.
Une étude quantitative a été réalisée sur sept sujets (P_1…P_7) atteints de maladie de Parkinson (60,5±4,3 ans), à partir d'une série d'acquisitions faites au CHU de Tours par le Dr Caroline Prunier. Pour chacun des sujets, une IRM pondérée en T_1 (128×128×128, 8mm^3) est disponible, ainsi que plusieurs acquisitions TEMP, recalées et redimensionnées par le clinicien sur l'IRM de référence, utilisant deux traceurs différents, l'un spécifique des récepteurs de dopamine en phase post-synaptique (^{123}I BZM), et l'autre (PE2I, non commercialisé en France), spécifique des transporteurs D2 en phase pré-synaptique. Pour ce dernier, les acquisitions ont été effectuées tous les quarts d'heure entre 15 et 65 minutes après injection. L'utilisation de ces deux traceurs est justifiée par le fait que l'atteinte en neurotransmission d'un patient MPI est uniquement pré-synaptique, tandis qu'elle est à la fois pré et post synaptique pour les syndromes parkinsoniens.
Les images, quantifiées manuellement par le Dr Marie-Odile Habert, ont permis de valider le résultat du diagnostic obtenu à l'aide des indices volumiques et quantitatifs calculés sur le striatum, en utilisant diverses zones non spécifiques (cortex, cervelet et cerveau complet) pour la normalisation. Les indices obtenus pour la fixation du PE2I ont été comparés aux données fournies par Prunier *et al.* [PRUNIER03], concernant un groupe de sujets sains et parkinsoniens en utilisant la méthode de Logan. Les 7 sujets présentés ici font partie du groupe des sujets parkinsoniens. Les auteurs utilisent une région d'intérêt tracée manuellement dans le cortex occipital comme région non spécifique. Les indices étudiés seront donc également normalisés par cette région. Toujours d'après les auteurs, la fixation rapide (15 à 20 minutes), puis la décroissance extrêmement rapide du PE2I rend intéressante l'étude des acquisitions correspondantes en temps précoce (à t+15)
Les résultats de la méthode proposée ici sont en complet accord avec ceux établis par Prunier *et al.* Les auteurs se sont particulièrement intéressés à l'asymétrie de fixation dans les putamens, liée à l'asymétrie de l'atteinte chez le sujet parkinsonien et d'autant plus importante que la pathologie est naissante. Les indices calculés révèlent une asymétrie absolue très faible chez 5 patients (Tableau 1-3), ayant tous plus de 4 ans d'évolution de la pathologie, les deux autres patients (P_1 et P_4) présentant une asymétrie plus forte avec une durée de pathologie plus courte.

Patient	Asymétrie de fixation dans les putamens
P_1	-0,258
P_2	-0,026
P_3	0,051
P_4	-0,405
P_5	0,024
P_6	0,026
P_7	0,165

Tableau 1-3 : asymétrie de fixation dans les putamens

Comme dans [PRUNIER03], P_1 et P_4 ont également une atteinte des putamens statistiquement supérieure ($p<0.005$) (Tableau 1-4)

Patients	Fix moy / moy CORTEX OC
P_1	2,21
P_2	1,51
P_3	1,50
P_4	2,47
P_5	1,53
P_6	1,99
P_7	1,54

Tableau 1-4 : fixation normalisée moyenne dans les putamens

Laulumaa *et al*. [LAULUMAA93] suggèrent une utilisation de l'imagerie TEMP à l'IBZM pour le diagnostic de la MPI à un stade précoce (patient pas encore sous traitement et pathologie encore latéralisée). Le cervelet est utilisé comme zone non spécifique et les auteurs observent une asymétrie de fixation chez les malades, avec une fixation plus élevée dans l'hémisphère contralatéral aux symptômes parkinsoniens. Le Dr Habert, quant à elle, a réalisé une étude quantitative sur les images utilisées ici en normalisant la fixation des zones spécifiques par la fixation dans une région d'intérêt tracée dans le cortex occipital. Les régions non spécifiques utilisées ici seront donc le cervelet et le cortex occipital. Là encore, une asymétrie de fixation significative est observée. De plus, une étude de la fixation moyenne normalisée montre que les têtes des noyaux caudés ont une fixation moins importante que les putamens, et que le cortex occipital semble une région plus discriminante que le cervelet pour l'IBZM).
Les volumes obtenus pour les TNC et les PU ont été comparés à ceux obtenus pour 7 sujets sains (extraits de l'étude précédente) appariés en âge avec les sujets parkinsoniens. Les indices volumiques et les indices d'asymétrie obtenus pour les sujets parkinsoniens ne sont pas statistiquement différents de ceux obtenus pour les sujets sains appariés en âge (les indices volumiques des sujets parkinsoniens sont inclus dans l'intervalle de confiance moyenne sains $\pm$ 2 écart-type sains).
Les premiers résultats présentés ici sont encourageants pour les applications cliniques étudiées, et ouvrent la voie vers de nouvelles perspectives de développement sur ce thème en collaboration avec de nombreux centres cliniques français. C'est avec ces mêmes centres que se développe un autre axe de recherche sur la quantification fonctionnelle, utilisant cette fois-ci le processus de fusion d'images développé dans le chapitre 1.

2. Fusion d'images anatomiques et fonctionnelles

Les développements entrepris à l'ERIM depuis 1997 concernant la fusion d'images anatomiques et fonctionnelles ont tout naturellement amené à rechercher des outils de synthèse d'information, permettant par exemple de présenter au clinicien sur une seule et même image l'ensemble des données disponibles. Dans cette optique, et pour poursuivre les premiers développements effectués lors de ma thèse (et

rappelés dans le paragraphe 2.2.2 du chapitre 1), j'ai été amené à co-encadrer la thèse de Julien Montagner effectuée dans le cadre du contrat RNTS « FusPark ».
Rappelons que le type de fusion envisagé ici concerne l'agrégation de N images anatomiques (typiquement IRM) et fonctionnelles (TEM, TEP) recalées. Les informations à combiner sont hétérogènes et les images ne sont pas informatives sur ce qu'elles ne sont pas censées représenter.

2.1. Modèles existants

A l'issue du processus de fusion, un ensemble de distributions de possibilité $\{\pi^i_T\}_{1\le i\le N}$ est calculé à partir des N images. Pour chaque tissu T, un opérateur de fusion (*cf.* 2.2.2 du chapitre 1) permet d'agréger les π^i_T en une distribution de possibilité π_T représentant l'appartenance de chaque voxel au tissu T.
Puisque les informations apportées par les deux types d'images sont de natures très différentes mais complémentaires, et puisque leur exploitation simultanée doit concourir à une meilleure compréhension du phénomène étudié, nous avons cherché dès 1997 à synthétiser une image représentant à la fois l'activité fonctionnelle et la précision anatomique, et surtout donnant accès à une quantification plus précise des phénomènes fonctionnels mis en jeu. Nous avons proposé dans ce but [COLIN97] [BARRA98-A] deux méthodes pour créer une image de synthèse à partir des distributions fusionnées π_T.

2.1.1. Synthèse non quantitative

Les degrés de possibilité sont ici considérés comme des pourcentages de volume partiel. Bien qu'aucun élément théorique ne vienne justifier cette assertion, il semble intuitivement satisfaisant de considérer que plus un voxel v a de possibilités d'appartenir à une classe de tissu T (*i.e.* plus $\pi_T(v)$ est élevé), plus le tissu T doit être présent dans ce voxel. Les pourcentages de volume partiel p_T des tissus T sont alors calculés après normalisation des degrés de possibilité :

$$p_T(v) = \frac{\pi_T(v)}{\sum_{Tissu} \pi_{Tissu}(v)}$$

A partir des valeurs d'activité fonctionnelle moyenne b_T issues d'étape de segmentation des images fonctionnelles, l'intensité de chaque voxel v de l'image de synthèse $SYNT$ est alors calculée par :

$$SYNT(v) = \sum_{Tissu} b_{Tissu}\, p_{Tissu}(v).$$

Cette méthode est simple mais très dépendante des valeurs b_i [COLIN97], qui ne représentent en outre que très approximativement les activités des tissus.

2.1.2. Synthèse avec préservation de l'activité locale

La méthode précédente ne permet pas de préserver l'information quantitative issue de l'image fonctionnelle. Dans le cas N=2 (une image anatomique, une image fonctionnelle), nous avons alors souhaité utiliser l'image anatomique pour "redistribuer" l'activité observée sur l'image fonctionnelle. A chaque voxel V de l'image fonctionnelle (de basse résolution spatiale, noté macro-voxel dans la suite) sont associés plusieurs voxels de l'image anatomique (de haute résolution spatiale). En notant :

p_T^v : pourcentage de tissu T dans le voxel v (contenu dans le macro-voxel V courant) ;
b_T : activité moyenne du tissu T, issue par exemple de la caractérisation tissulaire;
q_T : correction apportée localement à b_T afin de préserver la quantification ;
A : activité globale du macro-voxel V ;
n : taille de V (en nombre de voxels anatomiques),

l'égalité suivante est vérifiée :

$$(\forall V)\ \sum_{Tissu}\sum_{v} p^v_{Tissu}\, b_{Tissu}\, q_{Tissu} = \sum_{Tissu} q_{Tissu} b_{Tissu} \sum_{v} p^v_{Tissu} = n^3 A \quad .$$

La quantité $b_{Tissu}\sum_v p_{Tissu}^v$ est interprétée comme l'apport du tissu T dans l'activité de V.

Plusieurs règles sont envisageables pour le calcul des coefficients q_T. Une solution simple consiste à les considérer tous égaux, mais dans ce cas l'activité des tissus est systématiquement augmentée en cas d'hyperactivité d'une classe. Pour pallier cet inconvénient, et dans le cas de l'étude d'une pathologie affectant un tissu G donné (MG par exemple dans le cas des pathologies neurodégénératives), l'activité issue des zones segmentées comme G est modifiée en priorité. Pour ce faire, les coefficients q_i sont fixés à 1, sauf q_G, donné par :

$$q_G = \frac{n^3 A - \sum_{T \neq G} q_T b_T \sum_v p_T^v}{b_G \sum_v p_G^v},$$

qui définit une nouvelle image localisant plus précisément les données fonctionnelles (Figure 1-27).

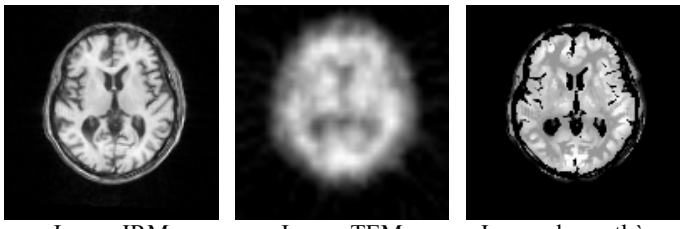


Image IRM Image TEM Image de synthèse

Figure 1-27 : image de synthèse

Quel que soit le modèle envisagé, les hypothèses sous-jacentes à la construction des images de synthèse stipulent que les degrés d'appartenance d'un voxel aux différentes classes de tissus peuvent être considérés comme des pourcentages de tissus dans l'élément de volume. En remarquant que les deux images peuvent être considérés comme des pavages discrets de l'espace 3D à des résolutions différentes, nous proposons d'affiner cette première approche en introduisant dans le calcul des activités de l'image de synthèse des éléments d'intersection discrète des voxels anatomiques et fonctionnels.

2.2. Méthode proposée

Nous raisonnons dans la suite dans le cas N=2. Les deux images sont mises en correspondance en gardant l'IRM comme référence, et en conservant le maximum de l'information fonctionnelle originale. Nous procédons pour cela à un changement de coordonnées discrètes, induit par les paramètres du recalage géométrique, mais permettant de repérer sans déformation les voxels de l'image TEM correspondant aux voxels de l'image IRM. Le faible rapport entre les deux résolutions spatiales implique, pour un changement de coordonnées précis, un passage par des calculs faisant intervenir les intersections entre les voxels des deux images. Les résolutions spatiales de ces deux images étant fort différentes, l'espace d'étude est divisé en deux pavages formés de voxels de tailles et d'orientations différentes, issus du recalage préalable. Le calcul des volumes de ces intersections doit permettre d'envisager une répartition plus objective et précise des activités fonctionnelles rapportées aux voxels anatomiques, et donc de préciser le modèle de formation de l'image de synthèse présenté dans le paragraphe précédent.

Dans le cadre de la thèse de Julien Montagner, nous proposons d'étudier ce problème d'intersection en considérant les similitudes qu'il présente avec le problème mathématique du « changement de coordonnées discrètes » [MONTAGNER02]. La relation entre les deux pavages 3D de l'espace est effectuée par rapport aux valeurs des volumes d'intersection entre les couples de voxels morphologique et fonctionnel. Ces volumes sont des polyèdres (Figure 1-28), calculés en utilisant un algorithme d'intersection de cubes efficace exploitant les symétries du cube pour déterminer des formules analytiques d'intersection [REVEILLES01].

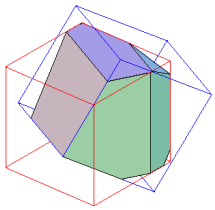


Figure 1-28 : polyèdre d'intersection entre deux voxels morphologique et fonctionnel

2.2.1. Construction des systèmes de voxels

La grille la plus fine (morphologique) est prise comme référence spatiale. Ses voxels sont simplement modélisés par des cubes unité, dont les faces sont prises par convention parallèles aux plans définis par la base canonique.

L'autre pavage du plan, issu de l'image fonctionnelle, est en position quelconque dans le repère de la grille morphologique (recalage). Ses voxels de plus grande dimension doivent alors être modélisés comme des cubes à coordonnées réelles dépendant des paramètres du recalage.

La taille des images originales (de l'ordre de 256^3 voxels pour l'image morphologique et 128^3 voxels pour l'image fonctionnelle) soulève ici le problème du temps de calcul de l'ensemble des intersections possibles : même avec un algorithme efficace de calcul d'intersection, le nombre d'itérations pour l'image complète reste important et doit être réduit. C'est la raison pour laquelle la modélisation des voxels proposée ici prend en compte les caractéristiques des treillis discrets, et en particulier les périodicités rencontrées sous l'hypothèse que les lignes générant les arêtes des voxels ont des directions entières. Dans le plan, si Δ est une droite réelle contenant l'origine, une direction entière implique que Δ passe par un autre point entier *P(a,b)*. Tous les points *k(a,b), k∈Z* appartiennent à Δ et les intersections entre la grille générée par les vecteurs unités et [*O*,*P*] sont alors répétées le long de Δ. L'intersection avec un carré unité dont un sommet est en *(i,j)* est alors identique à l'intersection avec le carré unité de sommet *(i+a,j+b)* (Figure 1-29).

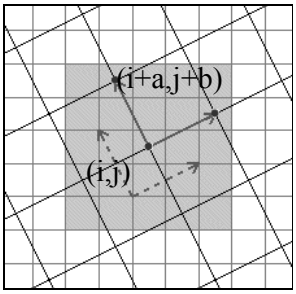


Figure 1-29 : prise en compte des périodicités pour le calcul d'intersection

Dans le cas 3D, de telles propriétés vont permettre de réduire de façon substantielle le nombre d'itérations dans le calcul des intersections, en calculant simplement des approximations entières des vecteurs générateurs V_1, V_2 et V_3 du système de voxels fonctionnels.

2.2.2. Approximation des directions

- Première direction

Soit $u = (r, s, t)$ un vecteur réel de norme 1. Nous recherchons les approximations diophantiennes de u, *i.e.* les vecteurs entiers αn qui approchent u d'aussi près que possible. Sans restreindre la généralité du propos, nous supposons dans la suite que r est la plus grande des trois composantes de u. Les intersections de la droite λu avec les plans $x = k$, $\lambda \in R$ et $k \in Z$, sont données par $(k, ks/r, kt/r)$. En notant $\sigma=s/r$ et $\tau=t/r$, le problème se ramène à l'étude des restes des suites modulaires $k\sigma \bmod 1$ $(\{k\sigma\})$ et $k\tau \bmod 1$ $(\{k\tau\})$. On montre alors que les valeurs de k satisfaisant le problème sont celles pour lesquelles les couples $(\{k\sigma\}, \{k\tau\})$ sont proches de l'un des quatre sommets du carré $[0,1]^2$, et sont déterminées en minimisant la suite $approx_k = Min(\max(\{k\sigma\},\{k\tau\}), \max(\{k\sigma\},1-\{k\tau\}), \max(1-\{k\sigma\},\{k\tau\}), \max(1-\{k\sigma\},1-\{k\tau\}))$. Sur la Figure 1-30 par exemple, les d_i sont les distances du point d'intersection (en rouge) de la droite λu avec

les carrés unités aux sommets de ces mêmes carrés. Pour l'ensemble des plans entiers $x = k$, le sommet du carré unité qui minimise d_i pour toute valeur de k est retenu comme extrémité de l'approximation diophantienne de la direction V_1.

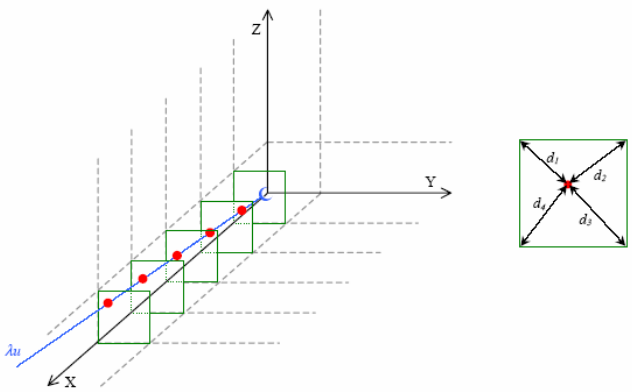


Figure 1-30 : recherche itérative de la première approximation diophantienne

En pratique, une centaine d'itérations (k=100) permet d'obtenir une approximation entière précise de la première direction.

- Seconde direction

L'approximation diophantienne de V_2 est déterminée dans le plan dont V_1 est le vecteur normal. Pour ce faire, une base entière de ce plan est tout d'abord calculée en utilisant l'algorithme d'Euclide-Blankinship 3D [14], puis l'approximation entière de V_2 dans ce plan est effectué en remarquant que le support de V_2 a une direction réelle qui peut être encadrée par son entonnoir de Klein [MOUSSAFIR92]. La pente α de cette direction est plus précisément développée en fraction continue $(a_0, a_1, a_2 \ldots)$. Les suites (p_k) et (q_k) définies à partir de la suite (a_k) par :

$$\begin{cases} p_0 = 0, p_1 = 1, p_{k+2} = p_k + a_k p_{k+1} \\ q_0 = 1, q_1 = 0, p_{k+2} = q_k + a_k q_{k+1} \end{cases}$$

permettent d'associer à la fraction p_k/q_k le point entier du plan $e_k = (p_k, q_k)$. La suite des (p_k/q_k) convergeant vers α, les points e_k sont de plus en plus proches de la droite de pente α, direction de V_2 (Figure 1-31)

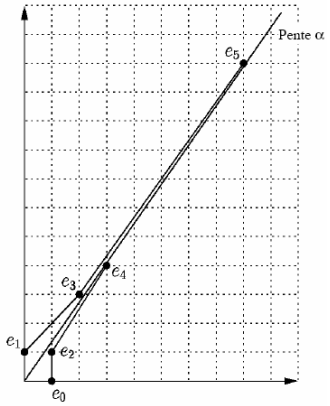


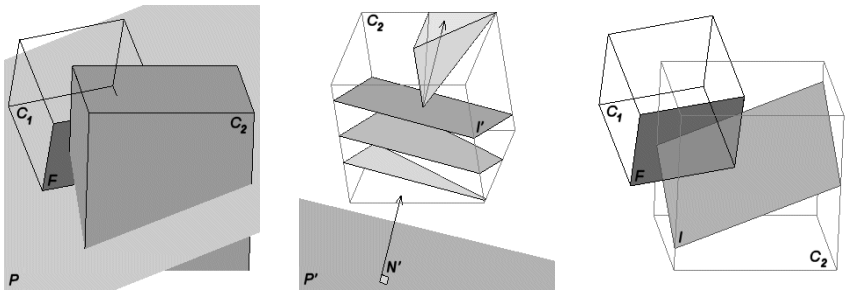
Figure 1-31 : interprétation graphique du développement en fraction continue

- Troisième direction

La troisième direction est enfin obtenue par un simple produit vectoriel des deux premières approximations diophantiennes.

2.2.3. Algorithme

Une fois les deux systèmes de voxels entiers calculés, l'algorithme général s'intéresse à l'intersection d'un couple de voxels des deux pavages. La méthode consiste à parcourir itérativement l'ensemble des 12 faces de ces deux cubes C_1 et C_2 (Figure 1-32-a). A chaque étape, le plan P support de la face courante F d'un voxel C_1 coupe C_2 selon un polygone I donné. L'intersection plane de ce polygone avec F fournit une des faces du volume recherché (Figure 1-32-c). Le polygone d'intersection I est déterminé par des formules analytiques [REVEILLES01]. Ces dernières sont établies en remarquant que si la direction normale à un plan P' est contenue dans un cône donné, le déplacement de P' suivant cette normale (et donc la distance de P' au centre de l'autre cube) fait varier son intersection I' avec le cube de manière connue (Figure 1-32-b). Il est possible de faire correspondre un tel plan P' à tout plan P support d'une face par une transformation T exploitant le groupe des symétries du cube (Figure 1-33). Le polygone I est alors retrouvé par application de T^{-1} au polygone I'. L'utilisation de ces formules pré-établies allège considérablement la première étape de l'algorithme.



(a) (b) (c)

Figure 1-32 : schéma d'intersection de deux voxels

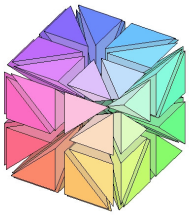


Figure 1-33 : groupe des symétries du cube

La seconde étape détermine la forme finale d'une face du volume recherché en un temps linéaire par rapport au nombre d'arêtes des deux polygones I et F [O'Rourke 98].

L'utilisation des périodicités des deux motifs de pavages permet enfin, à partir d'un nombre restreint de calculs d'intersections, de traiter le volume dans son ensemble [MONTAGNER03].

2.2.4. Quantification et synthèse de l'image

Une fois les intersections calculées, une quantification précise des activités fonctionnelles peut être effectuée, et un support visuel de cette quantification proposé. Pour ce faire, une activité arbitraire ou calculée à partir d'une information image est associée à chacune des classes anatomiques (matières ou structures). L'information morphologique est ainsi préservée dans les zones de transition entre ces différentes valeurs de base. L'information fonctionnelle qui doit apparaître dans l'image de synthèse concerne les variations locales d'activité dans chacune des zones anatomiques plutôt que la distribution globale. Cette information est injectée dans l'image synthétique selon le processus suivant : soit δ_v l'activité issue de l'image fonctionnelle originale *TEM* affectée au voxel v. Cette valeur est fonction des données portées par les macro-voxels V de l'image *TEM* présentant une intersection non nulle avec v :

$$\delta_v = \sum_{V \cap v} \rho(v, V) \cdot TEM(V)$$

Si les voxels de l'image anatomique de référence sont représentés comme les cubes précédemment décrits, le volume d'intersection $\rho(v,V)$ entre les voxels v et V appartient à [0,1]. δ_v apparaît donc comme une moyenne pondérée d'activités, et est elle même située dans le domaine des activités fonctionnelles. Finalement, la valeur portée par un voxel v de l'image de synthèse présentant une information fonctionnelle localisée est donnée par :

$$SYNTH(v) = \frac{\sum \pi_T(v) \cdot b_T + \delta_v}{\sum \pi_T(v) + \sum_{V \cap v} \rho(v,V)}$$

Bien sur, cette image n'a d'intérêt que visuellement, elle ne permet pas de quantifier plus précisément les activités fonctionnelles. Par contre, la méthode de génération de cette image donne accès pour chaque voxel v à une quantification prenant en compte les différences de résolution entre les deux images, qui est elle d'un grand intérêt pour le clinicien.

2.3. Premiers résultats

2.3.1. Contexte de l'étude

Le travail présenté dans cette partie s'intègre au projet « Fuspark » du Réseau National de recherche et d'innovation Technologies pour la Santé 2002, dont l'objectif général est de fournir des outils pour l'aide au diagnostic de la maladie de Parkinson et des syndromes parkinsoniens. Le protocole proposé utilise le couple IRM/TEM (imagerie d'un ligand du transporteur dopaminergique) pour mettre en évidence des défauts de concentration en dopamine localisés, représentatifs de la dégénérescence neuronale responsable de ces pathologies. Le traitement des images vise à l'obtention d'indices numériques de quantification, représentatifs de ce type de neurotransmission.

Une première étude menée avec le concours des 6 centres cliniques partenaires du projet a pour but la mise en place d'une base de données de référence pour les syndromes parkinsoniens. Elle comporte dans un premier temps les images de 30 volontaires sains et de 60 patients répartis en trois groupes correspondant à des pathologies identifiées (maladie de Parkinson idiopathique, atrophie multi système, paralysie supra nucléaire progressive), et un quatrième groupe de diagnostic atypique à classer parmi les trois autres en fonction de l'analyse quantitative. Dans ce projet, notre travail consiste notamment en la mise en place des outils informatiques de traitement des couples d'images IRM / TEM pour l'analyse quantitative. L'extraction des indices numériques représentatifs de l'activité dopaminergique dans les structures d'intérêt est réalisée par une fusion des informations anatomiques et fonctionnelles, et par l'exploitation de l'image de synthèse dont le processus de formation est décrit dans le paragraphe précédent.

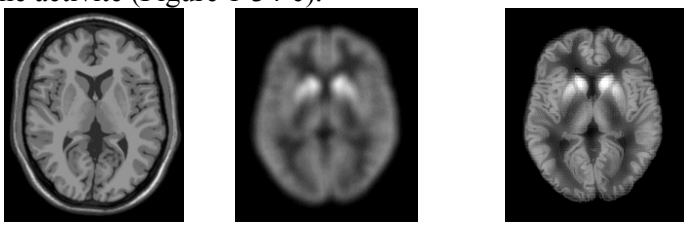
2.3.2. Validation sur fantômes

Le calcul des données géométriques dépend fortement du recalage entre les deux images, et la validité de l'information volume est très liée à la qualité de cette étape. La validation du procédé de synthèse d'image a été réalisée dans un premier temps sur fantôme numérique. Des données fonctionnelles sont calculées à partir des données anatomiques, et les deux jeux de données originaux sont donc alignés par définition. Une déformation rigide et un changement d'échelle sont appliqués à une des deux images afin de simuler les différences de position, d'orientation et de résolution spatiale. Dans ce contexte, les paramètres de recalage sont maîtrisés, car obtenus directement par transformation inverse.

- Fantôme d'étude de la perfusion

Un premier fantôme numérique a été conçu pour tester la synthèse d'une image portant une information fonctionnelle de perfusion cérébrale localisée. Des cartes d'appartenance aux tissus d'intérêt (LCS, MB, MG) sont d'une part obtenues par classification floue du fantôme anatomique de l'Université McGill [KWAN96]. Chaque voxel se voit d'autre part attribuer une combinaison d'activités fonctionnelles *a priori* (0 pour le LCS, *a* pour la MB et 4*a* pour la MG, valeurs correspondant à des niveaux relatifs de perfusion

tels qu'avalisés par des experts), pondérées par les coefficients d'appartenance aux différents tissus. L'image résultante subit ensuite un changement d'échelle, avant l'application d'un filtre moyenneur modélisant de manière très simple les effets de diffusions inhérents aux images TEMP. Les images du fantôme anatomique original sont des tableaux tridimensionnels de 60 coupes de 184x220 pixels, et les données sont supposées isotropes (Figure 1-34-a). L'image fonctionnelle simulée a subi une réduction par un facteur 2/3. Les données correspondantes se présentent sous la forme d'un tableau de données isotropes de 40 coupes de 120x144 pixels (Figure 1-34-b). Aucune rotation ou translation n'ont été appliquées pour cette première application. Les activités affectées à la région du striatum ont été doublées (8*a*). Bien que cela ne corresponde à aucune réalité clinique, cette différence permet de tester la préservation locale d'une activité (Figure 1-34-c).



a- Image IRM b- simulation TEM c- image de synthèse

Figure 1-34 : fantôme numérique de McGill pour la synthèse d'image

Les classes anatomiques d'intérêt se distinguent nettement sur l'image de synthèse. Les seules variations locales présentes dans les données fonctionnelles sont dues à la diffusion et à l'activité doublée de la zone du striatum. On retrouve ces variations locales sur l'image finale, sans que cette information vienne interférer avec l'information anatomique. Le contraste entre la valeur de base (activité moyenne) du striatum et celle de la matière blanche environnante entraîne une atténuation visuelle de l'activité 8*a* qui a été diffusée dans la MB. Les variations locales sont surtout visibles dans le striatum en lui-même.

- Fantôme d'étude de la neurotransmission dopaminergique

Un second fantôme numérique a permis la synthèse d'une image portant une information fonctionnelle de neurotransmission localisée par rapport aux structures cérébrales. Les données fonctionnelles du fantôme ont été obtenues par simulation de Monte-Carlo [ELFAKHRI01]. A partir d'un niveau de fixation de traceur supposé dans chacune des zones définies par le fantôme anatomique de Zubal [ZUBAL96] (Figure 1-35-a), les données fonctionnelles sont obtenues par simulation d'émission de photons gamma, et de modification de leur trajectoire par des cartes d'atténuation définies à partir de ce même fantôme (Figure 1-35-b). Les images du fantôme anatomique sont des tableaux 3D de 128 coupes de 256x256 pixels, et les données sont supposées isotropes. Les images du fantôme fonctionnel sont également isotropes, et se présentent sous la forme de tableaux de 60 coupes de 128x128 pixels. Le facteur d'échelle entre les deux images est de 2. Une déformation rigide (rotation seulement) a été appliquée à l'image anatomique, afin de préserver l'information fonctionnelle originale pour une future étude quantitative. L'image simulée est obtenue à l'aide de l'ensemble de ces données (Figure 1-35-c).

a- fantôme numérique de Zubal b- Carte d'atténuation c- Image fonctionnelle simulée d- Image de synthèse

Figure 1-35 : fantôme numérique de Zubal pour la synthèse d'image

Les variations locales de l'activité fonctionnelle de l'image de neurotransmission sont issues de la simulation, et donc plus indépendantes de l'image de départ. On retrouve ces variations sur l'image

synthétisée (Figure 1-35-d), surtout dans les zones d'activité importante (striatum), mais au détriment de l'information anatomique. Les petites variations entre les valeurs d'étiquetage des noyaux gris centraux du fantôme anatomique sont atténuées par le mélange avec les valeurs élevées de l'activité fonctionnelle. Les variations d'amplitude moindre sont moins perceptibles, du fait de la taille des structures anatomiques et du faible contraste entre elles. Dans une zone de taille plus importante (cervelet) à niveau de gris constant, ce problème d'atténuation n'apparaît pas.

Conclusion

Les deux approches présentées dans ce chapitre, et validées par ailleurs, ont pour but ultime de préciser des informations de type fonctionnel dans des zones d'intérêt cérébrales, et doivent donc permettre de proposer en clinique une information fonctionnelle quantitative localisée. Cet objectif peut être atteint soit en segmentant les structures d'intérêt sur une image morphologique, et en quantifiant les activités dans ces zones, soit en redistribuant les activités fonctionnelles dans les voxels morphologiques par des approches issues de la géométrie discrète. Les perspectives de ces deux techniques sont nombreuses.

Pour la première méthode, il s'agira par exemple de rechercher des scénarii adaptés à l'étude d'autres pathologies nécessitant un couple de données anatomiques et fonctionnelles, ou encore la recherche de paramètres quantitatifs pertinents pour ces mêmes études.

Pour la seconde approche, les premiers résultats sont prometteurs et les images de synthèse générées doivent maintenant être qualitativement validées par les partenaires cliniciens du RNTS. Elles sont le support visuel de la nouvelle répartition volumique des activités fonctionnelles proposée par la méthode développée ici : si elles ne présentent pas un intérêt clinique majeur de prime abord, elles donnent par contre accès à une nouvelle quantification informatique des données fonctionnelles, plus proche de la réalité des images de par l'intégration des pavages discrets dans le calcul. Cette quantification pourra quant à elle être validée de manière objective de plusieurs façons : par l'intermédiaire de fantômes numériques, par des comparaisons *a posteriori* de diagnostics cliniques ou encore par comparaison à des modèles biophysiques (cinétique et modélisation de distribution du transporteur de la dopamine). Dans le cadre de l'étude sur les syndromes parkinsoniens, les indices quantitatifs retenus devront refléter l'atteinte du versant pré-synaptique de la voie nigrostriée dopaminergique. Une fois validés, ces paramètres quantitatifs seront utilisés pour classer les patients du sous-groupe « syndrome parkinsonien atypique ». La vérification du bien-fondé de cette classification sera faite sur le suivi évolutif de ces patients. Si cette méthode a dans un premier temps pour objectif de démontrer et de valider l'intérêt du couple d'imagerie IRM/TEM pour délivrer des aides aux diagnostiques pertinentes et un suivi approprié dans le cadre de la maladie de Parkinson, son intérêt à plus long terme est de proposer une aide au clinicien pour toute étude de pathologie reposant sur une imagerie cérébrale quantitative localisée. Les partenaires impliqués dans l'étude proposent d'ailleurs déjà d'appliquer ces concepts dans de nombreux cas (épilepsie, Alzheimer), et le partenaire industriel de l'étude porte actuellement la méthode sur ses stations de travail.

Plus globalement, nous envisageons maintenant de combiner les deux types d'approches, en redistribuant simplement les activités fonctionnelles dans des structures d'intérêt préalablement segmentées (*e.g.* les striatum), et proposer ainsi un outil complet de quantification cérébrale localisée. Cet enjeu est d'importance, puisqu'il n'existe par exemple aujourd'hui aucune application clinique dont le diagnostic fasse appel à des indices quantitatifs localisés aux structures validés permettant une imagerie de routine de la neurotransmission.

Chapitre 4 - Stimulation magnétique transcrânienne

Seule la paresse fatigue le cerveau.
Louis Pauwels (L'apprentissage de la sérénité).

Introduction
Débutée en 1997, la collaboration avec Jean-Jacques Lemaire, neurochirurgien au CHU de Clermont-Ferrand, s'est concrétisée tout au long des sept dernières années par un certain nombre d'applications, dont une partie est décrite dans ce mémoire. Nous avons débuté en septembre 2003 une nouvelle collaboration concernant l'étude de la stimulation magnétique transcrânienne, dans le cadre d'une thèse CIFRE (Sébastien Luquet, Solusciences) que je suis amené à encadrer à temps plein sur le plan scientifique. Ce chapitre se veut une présentation succinte du sujet de thèse proposé. La stimulation magnétique transcrânienne est tout d'abord présentée, et les phénomènes physiques et biologiques mis en jeu sont ensuite décrits. Les objectifs que nous nous sommes fixés d'un point de vue théorique, applicatif et clinique, sont ensuite développés, et concernent une modélisation du phénomène mis en jeu dans la stimulation, à la fois d'un point de vue géométrique, biologique et physique, l'expérimentation prévue pour valider le modèle, et les retombées cliniques espérées. Les perspectives de ce travail sont enfin abordées, et concernent à court terme la réalisation d'un simulateur efficace du processus de stimulation en conditions données, à moyen terme la recherche des meilleurs sites de stimulation en fonction du matériel disponible et de la pathologie étudiée, et à long terme la création d'une plateforme informatisée d'aide à la stimulation magnétique.

1. La stimulation magnétique transcrânienne

1.1. Historique

La stimulation magnétique est une méthode permettant de stimuler des tissus excitables par un champ électrique, induit par les variations temporelles d'un champ magnétique extérieur. Outre les stimulations de la rétine et des nerfs, la stimulation magnétique transcrânienne (SMT) est une des applications les plus intéressantes et prometteuses, réalisée pour la première fois sur le cortex moteur de l'homme par Barker *et al.* en 1985 [BARKER85].

1.2. Principe

En stimulation magnétique transcrânienne, l'excitation extérieure est réalisée par l'intermédiaire de la circulation d'un courant intense pulsant *i(t)* dans une antenne *C* placée contre la tête du patient. Cette circulation génère par induction un champ magnétique *B*, donné par la loi de Biot et Savart :

$$B(r,t) = \frac{\mu_0}{4\pi} \oint_C \frac{dl(r') \wedge (r - r')}{|r - r'|^3},$$

la circulation étant calculée sur l'élément de longueur *dl* le long des spires de l'antenne *C*, avec la perméabilité du vide

$$\mu_0 = 4\pi.10^{-7} V.s / A.m .$$

Ce champ *B* est lui-même source d'un champ électrique *E*, induit dans les tissus. La forme de ce champ dépend de nombreux facteurs, parmi lesquels la forme de l'antenne *C*, la position et l'orientation de cette dernière par rapport aux zones de stimulation et les paramètres biophysiques des tissus cérébraux. Globalement, le champ électrique total dans un tissu donné est la somme d'une composante primaire E_1, induite par les variations de *B* et engendrant un vecteur densité de courant $J=\sigma E_1$, et secondaire E_2 créée aux interfaces des tissus par les changements de conductivité. En exprimant B à l'aide du potentiel

vecteur A ($B=rot(A)$), et en notant V le potentiel électrostatique généré par les accumulations de charges aux interfaces des tissus, le champ total E s'écrit :

$$E = E_1 + E_2 = -\frac{\partial A}{\partial t} + -\nabla V \qquad (1)$$

Ce champ affecte directement les neurones proches de la zone de contact de l'antenne, par l'intermédiaire d'un courant d'induction. Dans le cas d'une fibre supposée rectiligne, la composante axiale de ce courant a pour expression :

$$I_{ind} = \frac{E_i}{r_i},$$

où r_i est la résistance axiale de la fibre et E_i la composante axiale de E. La densité de courant transmembranaire supplémentaire induit par la stimulation est alors dérivée de la relation précédente, le long de la fibre repérée par son abscisse curviligne s :

$$I_{tr} = \frac{\partial}{\partial s}\left(\frac{E_i(s)}{r_i(s)}\right).$$

Les variations de diamètre, d'orientation de la fibre, ainsi que les gradients spatiaux du champ E vont varier ce courant $I_{tr.}$

1.3. Mécanismes biophysiques

La stimulation effective des tissus neuronaux est engendrée par E, qui induit une dépolarisation au niveau neuronal. Les mécanismes biophysiques régissant cette dépolarisation sont encore en grande partie inconnus, même si des études théoriques [ROTH90] ont démontré qu'une impulsion magnétique brève, inférieure à la milliseconde, pouvait générer un potentiel d'action sur une fibre axonale droite. Ces mêmes études [KAMITANI99] ont cependant également démontré que la SMT induisait une période d'inhibition de plusieurs millisecondes, observées à la fois en électromyographie et en inspection visuelle, et que les mécanismes d'action de la stimulation sont bien différents entre une fibre axonale droite (courant transmenbranaire induit par les gradients de E) et des neurones corticaux (le champ électrique induit est quasi uniforme sur le neurone en raison de la faible taille de ce dernier en regard de la taille de l'antenne [KAMITANI01]). De plus, les conditions d'expériences (forme de l'antenne, protocole de stimulation, localisation et orientation de l'antenne) conditionnent de manière critique les effets de la stimulation, et donc seule une interprétation qualitative des effets de la stimulation en fonction d'une expérience donnée est pertinente.

De manière plus générale, la question qui se pose lorsque l'on essaye de comprendre l'action de la SMT au niveau neuronal est la suivante : la stimulation reproduit-elle une physiologie normale du cerveau, ou agit-elle différemment et dépolarisant et activant des groupes de cellules différents dans une aire importante du cerveau ? La réponse à cette question passe d'une part par des études sur l'animal, et d'autre part en combinant la SMT avec des techniques d'exploration fonctionnelle (IRMf, MEG par exemple).

1.4. Applications

Depuis sa première application en 1985, la SMT a vu son champ d'applications s'élargir, depuis la recherche de fonctions cérébrales supérieures [KRINGS97][WASSERMAN97] à l'étude de syndromes parkinsoniens [PASCUAL94][YOUNG97], en passant par la psychiatrie [PRIDMORE99], l'épilepsie [NETZU97], l'étude de paralysies [RAPISARDA96] ou encore le traitement de problèmes visuels [HASHIMOTO95]. Globalement, la majorité des applications cliniques s'intéresse à une stimulation non-invasive des réseaux moteurs central et périphérique, des cortex visuel et préfrontaux, du centre du langage et du cervelet. Le but est à la fois pronostic, diagnostic, de monitoring et de soin.

2. Projet d'étude

Modéliser la stimulation magnétique transcrânienne se révèle d'une grande importance pour la recherche du meilleur point de stimulation, pour la compréhension des mécanismes mêmes de stimulation, pour l'interprétation des expériences et pour l'aide à la conception de nouveaux matériels d'instrumentation. La modélisation peut être vue sous deux aspects distincts et complémentaires : le calcul des champs électromagnétiques et des courants induits dans les tissus par la stimulation, et la réponse des neurones à l'accumulation de charges provoquée par ces champs sur leurs membranes. La suite de ce chapitre propose une esquisse du projet que nous souhaitons mener à bien dans le cadre de la thèse de Sébastien Luquet.

2.1. Calcul du champ et des courants induits dans les tissus

2.1.1. Modèle de tête

Le problème posé ici suppose connu un modèle de la tête, à la fois d'un point de vue géométrique et biophysique. Les premiers modèles sont apparus à la fin des années soixante, pour le calcul du champ électrique intra crânien, avec dans la plupart des cas une représentation géométrique simple en sphères concentriques, chaque sphère représentant un tissu spécifique aux propriétés biophysiques constantes et données [RUSH69]. Simple à mettre en œuvre, ce modèle est cependant peu réaliste, et ne permet d'expliquer que des variations grossières du champ *E*. Certains auteurs ont alors proposé des modèles plus complexes, à partir par exemple d'une discrétisation en triangles des différentes surfaces de séparation entre tissus cérébraux [NUNEZ90], où d'un modèle numérique de tête basé sur les acquisitions du Visible Human Project [GABRIEL01].

Nous souhaitons ici exploiter la méthode de segmentation développée par ailleurs (*cf.* Chapitre 1 - 2.1) pour proposer un modèle réaliste de distribution des tissus cérébraux. Plus précisément, à partir des cartes des tissus d'intérêt (matière grise, liquide cérébrospinal des espaces sous-arachnoïdiens, os, scalp), nous proposons de définir un modèle géométrique en couches, spécifique à chaque patient puisque généré à partir de la segmentation de l'IRM acquise, et réaliste.

Ce modèle géométrique va servir de base à une modélisation du patient, à condition d'indiquer pour chaque tissu un ensemble de paramètres biophysiques pertinents pour l'étude de la stimulation magnétique, en particulier la conductivité, la résistivité, la perméabilité magnétique de ces tissus ou encore la densité et l'orientation locales des neurones. Ces grandeurs varient bien sûr selon les individus, et à l'intérieur même d'un tissu donné en raison par exemple de son anisotropie (matière grise). Les mesures effectuées dans la littérature comportent des facteurs de variations parfois importants, comme le montre par exemple le Tableau 1-5 sur le cas particulier de la conductivité des tissus normalisées par rapport à celles du cerveau, mesurée à basse fréquence.

Référence / Tissu	[GEDDES67]	[HAUEISEN95]	[MARINO91]	[NUNEZ81]	[RUSH69]
Scalp	230	230		-	222
Crane	16000	16000	8000	20000	17760
LCS	64	65	65	64	-
cortex	222	300	200	350	222

Tableau 1-5 : exemples de valeurs de conductivité pour différents tissus cérébraux

Une étude bibliographique poussée doit donc être effectuée sur ce domaine, pour disposer d'un modèle valide d'un point de vue biophysique. Une étude aux dimensions doit également être effectuée, pour déterminer l'influence de l'ensemble des paramètres du modèle en fonction des grandeurs caractéristiques mises en jeu. Enfin, pour ce qui concerne les paramètres d'orientation et de densité locale des cellules neuronales, il pourra être fait appel à la fois à des données biologiques connues et à des données externes (imagerie en tenseurs de diffusion pour l'orientation, imagerie de fixation aux benzodiazépines ou

concentration en N-acetylspartate pour la densité, myélo et cyto architecture pour la constitution des faisceaux de substance blanche et des couches de cellules).

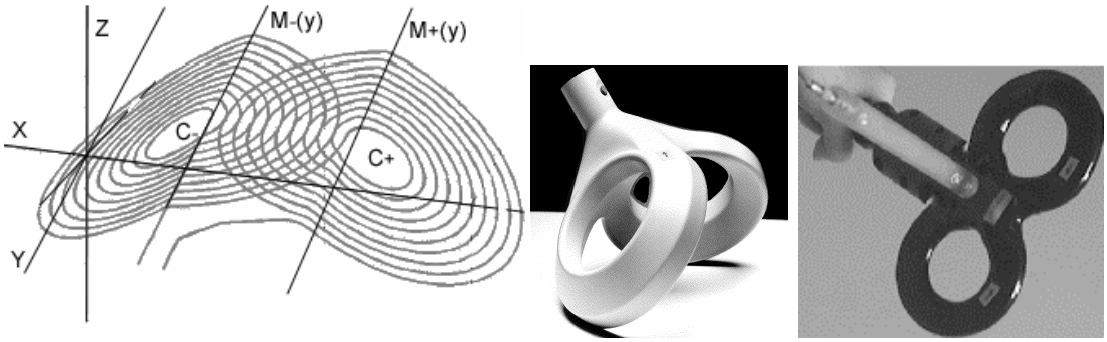
2.1.2. Dispositif de stimulation

L'antenne de stimulation dont nous disposons est le modèle MagPro X100 (Medtronic A/S^{TM}). Il s'agit d'une configuration géométrique en 8, et le dispositif propose quatre types de stimulations paramétriques : monophasique (stimulus simple), sinusoïdale (stimulations répétées de haute fréquence), sinusoïdale redressée et par salves.

Nous cherchons ici à modéliser l'antenne de ce dispositif, mais souhaitons également pouvoir généraliser ce modèle à d'autres types d'antennes existant sur le marché (bobine simple, antenne double cône). La modélisation retenue place donc le centre de l'antenne en $M=(x_0, y_0, z_0)$ et le dispositif est orienté dans le plan XY, le grand axe de symétrie du 8 étant selon X, et les centres des boucles du 8 étant en $C_{\pm}=M \pm (x_c,0,0)$. Les boucles du 8, de rayons interne (resp. externe) r_i (resp. r_e) sont composées de n spires et ont un degré de liberté angulaire θ_a autour des lignes $M_{\pm}(y)=(x_0 \pm r_i, y, z_0)$. La forme analytique des points $R(\theta)$ de ces boucles dans le plan XY est donc en polaire :

$$R(\theta) = (r_i,0,0) + \frac{r_e - r_i}{2\pi n}.|\theta - \theta_i|(\cos\theta, \sin\theta, 0) + C_{\pm},$$

où θ_i est la phase initiale. Les points pour lesquels $R(\theta)$ est en dehors de l'intervalle délimité par les lignes $[M_-(y), M_+(y)]$ sont transformés par rotation d'angle θ_a, de façon à obtenir une représentation réaliste et générique d'antennes de stimulation (Figure 1-36).



Modèle géométrique Antenne double cône Antenne en 8

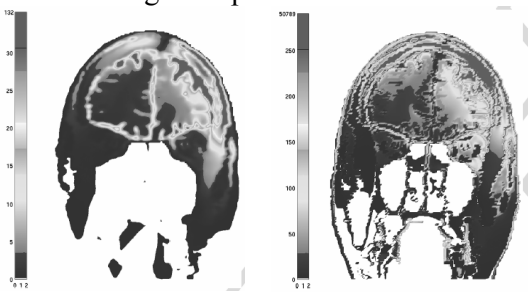
Figure 1-36 : modélisation et exemples d'antennes de stimulation

2.1.3. Modèle numérique

Différentes approches sont envisagées dans la littérature pour la recherche du champ E et/ou de la densité de courant J générés au niveau cortical par la stimulation magnétique. Le principe de base est une résolution des équations de Maxwell ou de l'équation (1) sur un domaine simple (sphérique [EATON92], cylindrique [ESSELLE95], infini [GRANDORI91]) ou complexe comme ceux précédemment modélisés, à l'aide de méthodes aux éléments finis [MIRANDA03], aux éléments frontière [DAVEY03], par la méthode de calcul d'impédance 3D [NADEEM03] ou par des calculs analytiques dans des configurations particulières [BRANSTON91] [GRANDORI91]. Les schémas numériques utilisés dans ces méthodes exploitent pour la plupart le fait que le champ électrique induit par une impulsion typique de SMT a un spectre en fréquence d'environ 10kHz : l'approximation quasi statique est alors justifiée pour la plupart des tissus biologiques [ROTH90], et les délais de propagation, les effets de peau et capacitifs peuvent donc être négligés. Peu d'études ont recherché les effets de l'inhomogénéité des paramètres biophysiques sur la forme de E et de J, et les principaux résultats concernent essentiellement l'observation d'un maximum du champ électrique dans les zones de faible conductivité électrique [CERRI95].

Une première approche de la bibliographie sur ce sujet nous a amené à considérer avec intérêt la méthode employée par Nadeem *et al.* [NADEEM03]. Les auteurs calculent une forme paramétrique du champ magnétique B généré par l'antenne à l'aide de la loi de Biot et Savart, en fonction des paramètres du

modèle géométrique d'antenne précédent, et en déduisent le champ électrique *E* et la densité de courant *J* dans un modèle 3D réaliste de tête (Brooks Air Force Laboratory, Texas) à l'aide de la méthode de l'impédance 3D (Figure 1-37). Pour ce faire, les paramètres biophysiques des tissus sont considérés constants et isotropes dans chaque voxel. Ces derniers sont alors vus comme nœuds d'un réseau d'impédances 3D sur lequel la loi des nœuds de Kirchoff est appliquée, les boucles de circuits considérées étant alimentées par le courant d'induction généré par *B*.



Variations de $|J|$ Variations de $|E|$

Coupe y=0.075 du modèle

Figure 1-37 : exemple de résultats obtenus par la méthode d'impédance 3D ([NADEEM03])

2.2. Confrontation à l'expérimentation

Quelle que soit la méthode retenue, il est indispensable de valider la phase de modélisation des distributions de *J* et *E*.

2.2.1. Etat de l'art

Si quelques études ont été réalisées sur fantôme pour mesurer les courants et le champ électrique générés par la stimulation magnétique [COHEN89] [MACCABEE90], une seule étude a été à notre connaissance effectuée *in vivo* sur l'homme [WAGNER04]. Cette dernière concernait un patient épileptique pour lequel des électrodes avaient été implantées dans un but de monitoring, et concluait à des densités de courant variant fortement en fonction de la position et l'orientation de l'antenne de stimulation, donnant alors accès à de nouvelles pistes d'interprétation des processus biophysiques impliqués dans la SMT. Cependant, cette expérimentation, bien que novatrice, n'a pas à notre sens été réalisée avec du matériel dédié (électrodes) et un protocole de positionnement de l'électrode bien défini.

2.2.2. Méthode envisagée

Dans une perspective à moyen terme, nous envisageons de valider la modélisation précédemment effectuée à l'aide d'expérimentations. Si pour l'instant aucune démarche précise n'est envisagée, les grandes étapes d'une telle phase doivent à notre sens comporter :

- la définition d'un système de mesure qui permettra de calculer les champs magnétique et électrique induits à l'intérieur du crâne. Diverses solutions ont d'ores et déjà été envisagées, depuis le placement de sondes dans des fantômes de cerveau jusqu'au placement de sondes du même type de façon plus ou moins invasive chez des animaux, en passant par des mesures de champ dans des tissus biologiques de tumeurs cérébrales ou d'animaux sacrifiés.
- la comparaison des résultats du modèle avec des données issues de techniques traditionnelles (IRM fonctionnelle par exemple).

2.3. Perspectives à long terme

Bien que débutant, cette étude propose de nombreuses perspectives, tant au point de vue applicatif que théorique.

D'un point de vue méthodologique, nous envisageons d'appliquer des techniques de fusion de données en agrégeant des données issues de la simulation à des informations fonctionnelles (type signal ou image). Le but étant de synthétiser l'ensemble des informations disponibles, par exemple pour proposer un protocole efficace de stimulation pour un patient donné dans un contexte donné, en terme de choix optimal de position et d'orientation de l'antenne, et de forme d'ondes de stimulation.

D'un point de vue clinique, et outre les nombreuses applications potentielles que cette technique propose (localement en psychiatrie notamment), nous souhaiterions que cette thèse aboutisse à la mise en place d'une plateforme de stimulation magnétique transcânienne assistée par ordinateur. Aujourd'hui encore, le positionnement de l'antenne se fait manuellement, et la localisation du point de stimulation optimal est encore hasardeuse, et fondée sur la seule expérience du praticien et sur quelques repères anatomiques simples. Les développements précédents devraient nous permettre de proposer à terme une interface agrégeant sur un même écran la position de la tête du patient, l'antenne de stimulation, et le modèle de la carte de champ électrique *E* calculé, la visualisation simultanée de l'ensemble de ces informations devant fournir une aide au positionnement précis de l'antenne. Il est même envisageable, grâce à des collaborations par exemple avec l'école de mécanique de Clermont-Ferrand (IFMA, Institut Français de Mécanique Avancée), de piloter un bras soutenant l'antenne de stimulation.

Conclusion

L'utilisation de la stimulation magnétique transcrânienne apparaît aujourd'hui comme une perspective d'avenir pour l'étude de pathologies, à la fois d'un point de vue pronostic, diagnostic, de monitoring et de soin. Ce chapitre se veut un plan d'étude de cette technique sur trois ans, dont les objectifs sont à la fois théoriques, méthodologiques et applicatifs, et dont le but final est la mise en place d'une plateforme de stimulation assistée par ordinateur. Bien sur, tous ces développements s'inscrivent dans le cadre du début de la thèse de Sébastien Luquet, et à ce titre sont uniquement à visée prospective.

Conclusion

Un jour viendra où tu croiras que tout est fini. Ce sera le début de tout.
Louis L'Amour (Lonely on the Mountain).

Les concepts théoriques initialement proposés en 1997 à l'ERIM pour le traitement de données médicales multi sources, et que j'ai développés plus avant lors de mon travail de thèse, s'inscrivent maintenant comme un cadre de travail validé et suffisamment générique, sur lequel reposent désormais bon nombre d'applications. Cette première partie avait pour objectif tout d'abord de rappeler les grandes bases de la méthodologie, puis de proposer les applications effectivement réalisées et à venir. Les perspectives sont en effet nombreuses.

A court terme, il s'agit tout d'abord de finaliser les travaux entrepris avec le CHU de Bordeaux concernant l'étude de la démence de type Alzheimer. Si la sectorisation anatomique du cortex est validée, il reste à l'appliquer sur l'étude multimodale IRM/TEM et à comparer *a posteriori* les résultats obtenus avec les diagnostics cliniques déjà effectués. De même, l'étude sur la sclérose latérale amyotrophique doit être menée à son terme, le but étant maintenant d'observer la fixation des traceurs spécifiques des récepteurs D2 de la dopamine dans le striatum (et plus particulièrement dans les têtes de noyaux caudés) afin de différencier, si possible, les sujets sains des sujets atteints de la pathologie, et de différencier également les types raides et non raides.

A moyen et long terme, il s'agira pour nous d'identifier en relation avec les cliniciens d'autres voies d'application de ces méthodes (étude de pathologies neurodégénératives spécifiques par exemple), et de développer les scénarii de segmentation et de quantification adaptés. Les problématiques cliniques ne manquent pas ici, et l'émergence de nouvelles applications induira nécessairement la modélisation et l'intégration de nouvelles connaissances dans le processus de fusion. Il peut par exemple s'agir de nouvelles techniques d'acquisition IRM pour la mise en évidence de phénomènes physico-chimiques reliés à la pathologie étudiée, ou encore de contraintes de formes floues de structures pathologiques (tumeurs par exemple).
Ces mêmes cliniciens feront d'autre part office d'experts pour la validation et la mise en application des nouvelles images de synthèse générées à l'aide des outils de géométrie discrète. Le but ici sera d'une part de démontrer que l'image de synthèse fournit en une seule visualisation le même type d'informations que l'ensemble des images d'entrée, et d'autre part de proposer ces développements sous la forme d'une plateforme logicielle en relation avec les entreprises partenaires du contrat RNTS. Enfin, l'étude du positionnement des électrodes en stimulation cérébrale profonde, appliquée au traitement de la maladie de Parkinson, doit se concrétiser à moyenne échéance, par l'intermédiaire notamment de la thèse d'Alice Villéger. Il s'agit ici pour nous de consolider le repérage des noyaux sous-thalamiques dans l'image IRM, et de proposer une interface de neuronavigation intégrée à l'existant (collaboration avec la société BrainLab), et présentant les résultats de la segmentation.

Enfin, les projets axés sur la stimulation magnétique transcrânienne constituent des perspectives de développement à plus long terme. Si l'étude ne fait que débuter aujourd'hui, elle n'en reste pas moins prometteuse en terme de retombées fondamentales et cliniques. Nous envisageons ici de mener de front

des aspects méthodologiques (choix de la méthode de résolution du problème aux dérivées partielles, construction du modèle biophysique de tête), théoriques (modélisation des populations neuronales en relation avec la biologie, modélisation du phénomène induit sur les neurones corticaux par la stimulation magnétique) et applicatifs (définition de protocoles de plans d'expériences pour l'étude et la validation des modèles, intégration de l'outil à une plateforme de stimulation assistée par ordinateur, applications cliniques locales en psychiatrie).

D'une manière plus générale, le concept de fusion tel qu'il est envisagé ici laisse la porte ouverte à une multitude d'applications cliniques fondamentales, que ce soit dans le domaine cérébral, ou pour l'étude d'autres organes (étude tomodensitométrie/TEM des cancers de la petite cellule en imagerie thoracique, ou des affections osseuses par exemple). Par exemple, nous avons appliqué tout ou partie de ces concepts dans l'étude d'images IRM de cuisses, pour diverses applications cliniques (étude du métabolisme musculaire [BARRA96] [BARRA98-C] [MORIO00], étude du rapport muscle/graisse chez le sportif et chez le malade atteint de la maladie de Cushing [BARRA02-C], Etude des volumes musculaires et graisseux suite à la reconstruction du ligament croisé antérieur [REBAI01-A] [REBAI01-B] [REBAI02], recherche la perte de masse musculaire due au vieillissement, suivi de la fonte musculaire en rapport avec des apports nutritionnels par voie entérale ou parentérale, étude en cours dans le cadre d'un PHRC avec l'HCL hôpital Lyon sud).

SECONDE PARTIE :
ETUDE DES IMAGES DE PUCES A ADN

Introduction

Le concept de puce à ADN repose sur une technologie multidisciplinaire intégrant la biologie, la nanotechnologie, la chimie des acides nucléiques, l'analyse d'images et la bioinformatique. Grâce à cet outil, il est possible de mesurer le niveau d'expression de plusieurs milliers de gènes simultanément, et les applications dans un grand nombre de domaines, dont font par exemple partie la pharmacologie et la médecine, sont en plein essor (détermination des familles de gènes co-régulés,...). Le traitement informatique des images de puces est une étape importante dans la lecture des données. Il peut être effectué par de nombreux logiciels, qui fournissent une quantification de l'expression des gènes étudiés. Cependant il n'existe aucune réelle évaluation des différentes étapes de ce processus de quantification (prétraitements de l'image, détection des spots, quantification des activations, classification et regroupement des gènes co-régulés), en particulier en raison de l'absence de référence absolue. Il paraît pourtant essentiel de pouvoir qualifier les différentes méthodes utilisées, et de quantifier leur performances en fonction de divers paramètres d'acquisition et d'utilisation.

Mon projet de recherche au sein du Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS, Université Blaise Pascal de Clermont-Ferrand) s'articule depuis 2001 autour de cette problématique.

L'analyse biologique des premières biopuces repose sur l'étude du transcriptome (étude des ARN messagers) d'une cellule ou d'un organisme. Si de nouvelles applications pour les puces à ADN ont également été développées (biopuces pour le diagnostic, pour la génomique comparative, pour l'identification de régions d'ADN régulatrices après ImmunoPrécipitation de la Chromatine (ChIPchips)), seules les biopuces transcriptome sont effectivement abordées ici (*cf.* Annexe A).

Cette seconde partie se propose alors de développer un cadre de test et de validation des méthodes utilisées en routine dans l'analyse des images de biopuces, par l'intermédiaire de la création d'un simulateur informatique d'images de biopuces transcriptome. Le premier chapitre présente ainsi le simulateur d'images de biopuces que nous avons développé au laboratoire. Partant du constat que de nombreux facteurs biologiques, chimiques, mécaniques et algorithmiques introduisent des variabilités dans le signal recueilli, nous avons construit un modèle statistique paramétrique, reprenant l'ensemble de ces facteurs de variations, et permettant de simuler des images de biopuces pour lesquels tous les paramètres, géométriques et de signal sont connus.
Les chapitres 2 et 3 détaillent ensuite un ensemble de techniques d'analyse des images de biopuces, intégrant la quasi totalité des méthodes couramment utilisées en routine. Pour chacune des quatre grandes étapes d'analyse (pré traitement, segmentation, quantification, extraction de connaissances), ces chapitres dressent un état de l'art, proposent de nouvelles méthodes et valident les méthodes existantes et nouvelles par rapport à des images simulées qui servent de référence. En particulier, une nouvelle méthode de segmentation/quantification des spots est proposée dans le chapitre 2, fondée sur une triangulation itérative de l'image guidée par son contenu, et deux nouvelles méthodes d'extraction de connaissances sont décrites, l'une reposant sur une représentation continue des profils des gènes et particulièrement adaptée à l'étude de données temporelles, l'autre autorisant une recherche de réseaux de régulation et fondée sur une utilisation de la fusion de données.

Le but final étant enfin, à la lumière des analyses précédentes, de proposer un nouvel ensemble logiciel réalisant l'analyse complète des images de biopuces et intégrant au mieux les réponses aux différents problèmes soulevés précédemment.

Chapitre 1 - Le simulateur de puces à ADN

J'ai trouvé cette chose étonnante : on peut représenter
par les nombres toutes sortes de vérités.
Gottfried Leibniz.

Introduction

Selon l'application désirée et le modèle biologique étudié, plusieurs types de biopuces peuvent être mises en œuvre. Pour l'étude du transcriptome par exemple, les biopuces ADNc ou oligonucléotides sont utilisées, tandis que les biopuces à ADN génomique sont consacrées à la génomique comparative ou à l'identification de régions d'ADN régulatrices après ImmunoPrécipitation de la Chromatine (ChIPchips). Dans tous les cas, des matériels et logiciels dédiés permettent d'acquérir des données relatives à ces puces, sous la forme d'images, et de traiter ces informations pour en extraire des informations biologiques pertinentes.

Bien que les méthodes d'analyse d'images réalisées dans les logiciels dédiés soient quotidiennement utilisées pour extraire des interprétations biologiques des biopuces, aucune réelle validation de ces algorithmes n'a été à notre connaissance effectuée en fonction des différents facteurs de variations biologiques, mécaniques, chimiques et algorithmiques inhérents à la formation de ces données spécifiques. Cette validation paraît pourtant essentielle, puisque de la bonne manipulation des informations dépend l'interprétation des données biologiques sous-jacentes.

Ce chapitre pose les bases de la construction d'un simulateur d'images de biopuces transcriptome. Les différents facteurs de variations recensés influencent soit la géométrie de l'image de puce, soit le signal effectif capté par la chaîne d'acquisition, et sont modélisés à l'aide de lois statistiques paramétriques et reflétant au mieux la variabilité des données effectivement observées. La mise en œuvre de cet outil de simulation est alors présentée, sous la forme d'un logiciel autonome et d'une plateforme WEB.

Ce simulateur permettra dans les chapitres suivants de générer des images paramétriques qui seront utilisées pour une étude prospective de validation des méthodes d'analyse des images de puces.

1. Le besoin de validation

1.1. Facteurs de variations

La création d'une puce à ADN est un processus complexe impliquant de nombreuses étapes. Les variations introduites lors de ces différentes phases peuvent nuire à la détection des vraies différences d'expression des gènes étudiés. Les causes d'artefact du signal dans une image de puce sont nombreuses, et surviennent à chaque étage de la fabrication d'une puce.

1.1.1. Sélection et préparation des sondes et cibles

Le marquage par un fluorochrome nécessite une étape de transcription inverse, qui permet de revenir à l'ARN messager de la séquence nucléotidique du gène initial. Durant cette réaction, les marqueurs fluorescents sont attachés à l'ARN. Cette manipulation implique des processus biochimiques complexes dont l'efficacité est très dépendante de nombreux paramètres difficiles à contrôler. De plus, l'ARN manipulé est susceptible de se dégrader rapidement, entraînant par exemple un dépôt insuffisant de sondes sur la puce et donc une sous-estimation (voire une non détection) de l'expression du gène correspondant [YUE01]. Enfin, les étapes de marquage et d'amplification du matériel biologique peuvent varier de façon non seulement systématique, en fonction de la composition des échantillons et de la stabilité des marqueurs, mais aussi aléatoire.

1.1.2. Dépôt des matériels biologiques

La fabrication effective de la puce implique le dépôt des sondes sur la surface du substrat, en utilisant les aiguilles d'un robot. En général, ce dernier fonctionne avec plusieurs aiguilles en parallèle, et le dépôt des

sondes peut donc varier d'une aiguille à l'autre. De plus, l'efficacité d'une aiguille donnée peut varier au cours du processus de dépôt, en raison par exemple de causes mécaniques, induisant un différentiel dans les quantités déposées sur le substrat. Ce dernier (verre, silicium) influence enfin directement la détection des expressions de gènes, de par sa préparation avant le dépôt (application d'un film de polymère, qui se doit d'être le plus uniforme possible, et dont le rôle est d'assurer la liaison électrostatique des sondes) et de sa composition propre (un substrat en verre générera par exemple lors de l'acquisition une diffusion du laser).

1.1.3. Hybridation

L'efficacité du processus d'hybridation est fonction de paramètres expérimentaux tels que la température, la durée et le nombre de molécules sondes déposées sur la puce. Lors de cette étape, des hybridations non spécifiques sondes/cibles (*i.e.* l'appariement d'une sonde avec un brin qui ne lui est pas entièrement complémentaire) peuvent également se produire, participant au bruit de fond détecté lors de l'étape d'acquisition.

1.1.4. Acquisition de l'image

L'étape d'acquisition implique de nombreux appareils et traitements, chacun disposant de ses paramètres propres. Citons ici tout particulièrement le scanner, caractérisé entre autres par sa calibration, sa résolution ou sa sensibilité de détection, et d'une manière plus générale la chaîne d'acquisition dans son ensemble, influençant le signal détecté par les caractéristiques de la source lumineuse, du système optique de détection, par l'action des photomultiplicateurs ou encore la mécanique d'acquisition.

1.1.5. Traitement de l'image

Enfin, le paramétrage et la fiabilité des algorithmes de traitement des données acquises influence de manière directe l'interprétation biologique des résultats. Ainsi par exemple le choix des méthodes de filtrage et de normalisation des données, la méthode de segmentation des spots, l'estimation du fond ou le choix des indices de quantification sont à considérer.

1.2. Importance de la validation

Pour réduire au maximum les risques d'erreurs, plusieurs solutions sont envisageables. Il est tout d'abord possible de répliquer plusieurs fois les mêmes gènes sur la même puce, et de pratiquer une étude de reproductibilité sur les données [LEE00]. Cette technique implique cependant une perte de place sur la puce, et comporte de nombreux autres inconvénients (en raison de la technique d'acquisition, le niveau d'expression d'un même gène à différents emplacements n'est par exemple pas nécessairement identique). Les puces peuvent également être préparées « localement à façon », dans des conditions bien contrôlées, pour donner les meilleurs résultats possibles. Cependant, une légère fluctuation des conditions expérimentales peut produire des résultats inexploitables. Enfin, si quelques auteurs proposent des méthodes permettant de contrôler certaines étapes de la fabrication (*e.g.* [HESSNER03] pour la sélection des sondes, [FRANSSEN02] [HANDRAN01] pour l'étape d'hybridation, [BASARSKY00] [DORSEL99] [PICKETT02] [RAMDAS01] pour l'acquisition, [MACHL02] [SPEED02] pour l'étape de traitement informatique), aucune ne permet à notre connaissance de gérer l'ensemble du processus de création d'une puce.

De nombreux auteurs [BALAGURUN02] [KOTHAPALLI02] [LALUSH99] [MOODY02] [WIERLING02] indiquent d'une manière générale qu'il est très difficile d'évaluer les méthodes utilisées dans les logiciels et la fiabilité des données extraites. Balagurunathan *et al.* [BALAGURUN02] ajoutent que cet état de fait est dû au manque de connaissance sur les vrais niveaux d'expression des gènes. Tous ces auteurs, et d'autres encore [Le MEUR01] [YANG00], évaluent donc les différents algorithmes de traitement sur des données publiques ou à façon, et estiment de manière relative les performances des diverses méthodes, sans vraiment prendre en compte la nature et l'importance des différents facteurs de variation du signal. L'interprétation des données biologiques sous-jacentes peut donc parfois s'en trouver biaisée.

L'avancée prodigieuse que proposent les puces à ADN va de plus amener biologistes et bioinformaticiens à vouloir détecter des relations de plus en plus fines entre les groupes de gènes, qui pousseront à la limite de leur technologie ces puces et les méthodes d'analyse utilisées (rapport signal sur bruit plus faible,

significativité des niveaux d'expression plus difficile à interpréter,…). De nombreuses entreprises et laboratoires bioinformatiques (*e.g.* Nutec Sciences, Imaging Research, axon, Université de Stanford) sont d'ailleurs déjà en quête de telles réalisations.

Pour toutes ces raisons, il est donc essentiel de proposer une méthodologie de validation et de calibration des différents outils utilisés en analyse d'images de puces. A notre connaissance, aucune méthode de validation des méthodes réalisées et commercialisées pour l'analyse des puces à ADN n'a été proposée, en fonction des divers facteurs paramétrant l'image de la puce. La suite de ce chapitre propose une contribution dans ce sens.

2. Le simulateur d'images de puces à ADN

Quelques auteurs proposent une modélisation des images de puces à ADN. Wierling *et al*. [WIERLING02] et Balagurunathan *et al.* [BALAGURUN02] développent en particulier une simulation fondée sur des lois statistiques, mais certaines de leurs propositions (concernant en particulier le bruit de fond et les modèles physiques sous-jacents) ne sont pas à notre avis justifiés par rapport au processus physique de formation du signal effectif (*cf.* infra). Katajamma [KATAJAMMA03] développe pour sa part un modèle fondé sur une décomposition en couches du processus de formation de l'image (étages biochimique, d'hybridation, d'acquisition), chacun de ces étages étant modélisé par des lois statistiques variées. Lalush propose quant à lui [LALUSH99] une simulation fondée uniquement sur un modèle du signal à base de lois normales et géométrique fondée sur les champs de Markov. Enfin, la société Affymetrix offre la possibilité *via* un portail WEB (http://affycomp.jhsph.edu, [COPE03]) d'évaluer les mesures des expressions de gènes sur leurs puces à partir de données spécifiques à chacun (réplication, dilution…).
Tous les paramètres évoqués précédemment permettent de jouer sur la géométrie (résolution, pose des sondes,..) et sur le signal (bruits de diffusion, d'hybridation, instrumental, activités des spots) des images résultantes. Nous avons identifié et proposons ici une modélisation de l'ensemble de ces paramètres au sein d'un simulateur de puces à ADN, dont un diagramme de blocs est proposé dans la Figure 2-1.

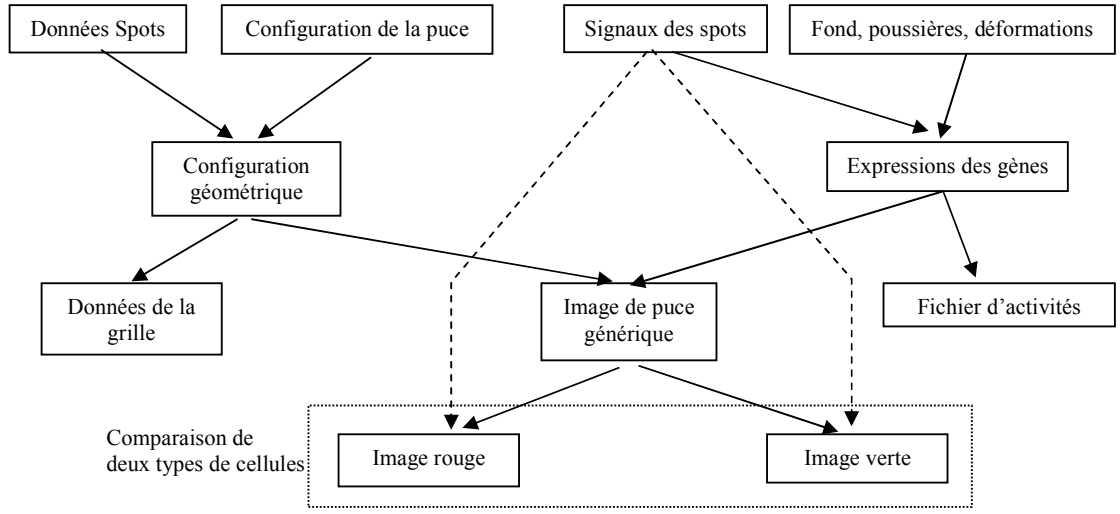


Figure 2-1 : diagramme de blocs du simulateur

2.1. Modélisation géométrique

La modélisation géométrique suit un processus hiérarchique et paramétrique illustré sur la Figure 2-2 : une puce est une collection de blocs, chacun d'entre eux étant défini par un ensemble de spots.

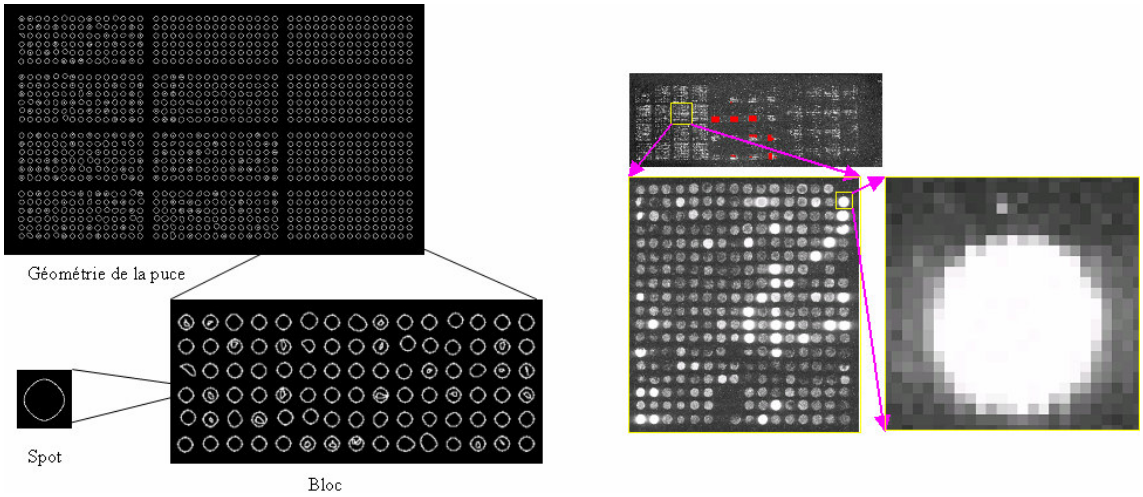


Figure 2-2 : processus hiérarchique de modélisation géométrique

2.1.1. Définition d'un spot

La frontière de chaque spot est modélisée par une B-spline cubique uniforme, définie par quatre points de contrôle *(P_0, ..., P_3)*, formant un polygone de contrôle $\mathcal{P}$:

$$P(t) = \sum_{i=0}^{3} P_i N_i^3(t), 0 \le t \le 1$$

où les N^3_i sont des polynômes de degré 3 définis par la relation de Cox-De Boor :

$$N_i^1(t) = \begin{cases} 1 \text{ si } t_i \le t \le t_{i+1} \\ 0 \text{ sinon} \end{cases}$$

$$N_i^k(t) = \frac{t - t_i}{t_{i+k-1} - t_i} N_i^{k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} N_{i+1}^{k-1}(t)$$

Les nœuds t_i sont uniformément distribués dans [0,1], permettant de relier le degré de la B-spline aux P_i. Connaissant la résolution du scanner R_s, il est alors possible de déterminer le rayon maximum des spots de la puce qui peuvent être détectés par le scanner modélisé, et donc de calculer le polygone de contrôle $\mathcal{P}$ en fonction de ce rayon. Par exemple, un spot circulaire de 100 µm acquis avec des pixels de 10 µm contient environ 80 pixels, ce qui est suffisant pour une analyse statistique des niveaux d'expression à l'intérieur des spots.

En raison de nombreux artefacts biologiques, physiques et mécaniques, les irrégularités dans les formes et les tailles de spots sont chose courante dans de vraies images de puces. La modélisation définie ci-dessus permet un calcul aisé et rapide de toutes ces déformations géométriques.

Les spots ont généralement une forme circulaire $\mathcal{C}(O,\rho)$, mais leur rayon n'est cependant pas constant sur toute la puce, en raison par exemple du comportement différent de deux aiguilles de dépôt distinctes. Puisque tous les spots d'un même bloc sont déposés par une même aiguille, le rayon de chaque spot d'un bloc donné peut être modifié par une variation de la position de tous les points de contrôle P_i. Cette modification de position est orientée par OP_i, O désignant le centre du spot théorique, et a pour norme un réel généré selon une loi normale $N(0,\sigma_r^2)$, σ_r^2 étant proportionnel au rayon initial du spot.

Cette forme circulaire est en fait souvent dégénérée sur de nombreux spots d'une image de puce :

- Des parties entières de spots peuvent disparaître durant le processus de fabrication de la puce, en raison par exemple du processus de lavage du substrat ou d'un phénomène de tension de surface au moment du séchage du substrat. A l'inverse, un dépôt trop important de sonde peut générer un spot beaucoup plus grand que prévu. Un spot peut donc en fait prendre une forme quelconque à partir de sa

forme circulaire théorique, ce que le simulateur restitue en permettant à un ou plusieurs points de contrôle P_i de se déplacer autour de leur position initiale selon une loi normale $N(0,\sigma_d^2 I_2)$. Dans la suite, ces spots seront nommés spots déformés.

- Lorsque l'aiguille de dépôt est trop proche du substrat, le matériel biologique déposé peut être éjecté du centre du spot théorique, laissant ce dernier vide. Connu sous le nom d'effet doughnut, ce phénomène génère un trou de signal au centre du spot, simulé en créant une B-spline cubique interne au spot, dont les points de contrôle Q_i sont positionnés en fonction de la frontière de ce spot. Plus précisément, la position de chaque point Q_i est uniformément distribuée dans un cercle de centre O et de rayon $F_d.\rho$, avec $0< F_d<1$.
- Le processus de séchage et la résolution limitée du scanner dégradent les frontières des spots. Ce phénomène est modélisé en ajoutant un bruit gaussien à la frontière des spots, dont la variance σ_e est fonction du rayon ρ du spot. Lorsque le spot est un doughnut, la B-spline interne est également dégradée.
- Le plan de dépôt suit théoriquement une grille régulière, définie par le robot. En parallèle de la modélisation géométrique de la puce est donc proposé un processus de génération de grille associée, pour lequel chaque case de la grille contient un et un seul spot. Pour modéliser les processus de vibrations mécaniques, induisant un non respect du plan de dépôt théorique, les spots peuvent se déplacer de leur position d'origine, par l'intermédiaire de la variation de la position de leur centre O selon une loi $N(0,\sigma_t^2 I_2)$, σ_t étant reliée à la distance inter-spot. Dans le cas d'un déplacement effectif, le spot concerné est dit spot déplacé.

La Figure 2-3 résume l'ensemble des géométries décrites ci-dessus.

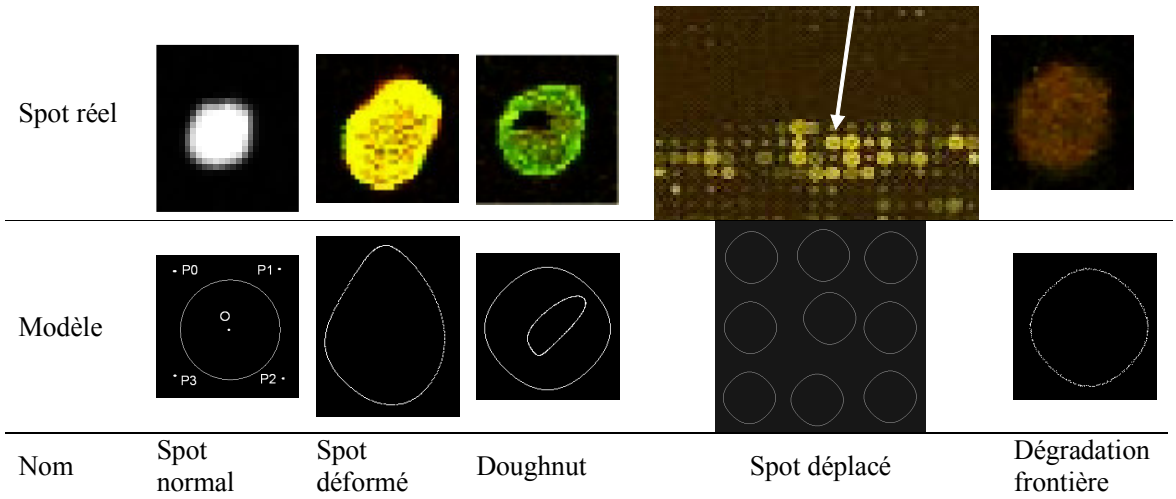


Figure 2-3 : exemples de géométries de spots

2.1.2. Structure de la puce

Les images de puce sont constituées d'un tableau de spots arrangés en sous-grilles appelées blocs, possédant le même nombre de spots en ligne et en colonne. Connaissant les espacements horizontal et vertical des spots dans un bloc, supposés constants, il est facile à partir de la modélisation précédente de construire de tels blocs. Cependant, l'espacement entre les lignes et les colonnes de spot d'un même bloc peut varier, en raison notamment de contraintes mécaniques imposées par le robot. Cette variation est modélisée en permettant à l'espacement horizontal (respectivement vertical) H_B (resp. V_B) entre les spots d'un même bloc de varier selon une loi normale $N(0,\sigma_b^2)$, dont la variance dépend de H_B (resp. V_B).

De même, les blocs d'une puce sont organisés selon un réseau régulier, perturbé là encore par des variations imposées par la mécanique d'acquisition. Cet espacement horizontal (resp. vertical) inégal des blocs est simulé par l'intermédiaire d'une loi normale $N(0,\sigma_m^2)$, dont la variance dépend de l'écartement théorique défini par le plan de dépôt horizontal H_M (resp. vertical V_M) entre les blocs.

Enfin, deux paramètres M_h et M_v, précisant les coordonnées de l'origine de la grille théorique dans le repère de l'image, permettent de préciser où se situe la puce sur l'image acquise.
La géométrie d'une puce à ADN est donc calculée à partir du jeu de paramètres précédemment introduit, auquel il faut ajouter des valeurs permettant de définir les pourcentage de spots déformés (P_{de}), doughnut (P_{do}) et déplacés (P_{di}) présents sur la puce. L'ensemble de ces paramètres est récapitulé dans le Tableau 2-1, et un exemple de modélisation géométrique est présenté dans la Figure 2-2.

Nom	Description	Fonction	Lié à	Définit
R_s	Résolution du scanner	Géométrie de la puce	Paramètre d'entrée	ρ, dimensions de la puce
σ_r	Variance de la loi normale N (0,σr²)	Variation de ρ	ρ	-
σ_d	Variance de la loi N(0,σd²I2)	Variation de géométrie du spot	ρ	-
F_d	Facteur d'échelle de ρ	Définition des points de contrôle du doughnut	ρ	-
σ_e	Variance de la loi N (0,σe²)	Dégradation de la frontière du spot	ρ	-
σ_t	Variance de la loi N (0,σt²)	Déplacement du spot	H_B, V_B	-
S_H,S_V	Nombres de spots en ligne et colonne dans un bloc	Description géométrique interne d'un bloc	Paramètre d'entrée	Dimensions de la puce
H_B, V_B	Espacements horizontal et vertical des spots dans un bloc	Description géométrique interne d'un bloc	Paramètre d'entrée	Dimensions de la puce
σ_b	Variance de la loi N (0,σb²)	Variation de l'espacement des lignes et colonnes dans un bloc	H_B, V_B	-
B_H,B_V	Nombres de blocs en ligne et colonne dans une puce	Géométrie de la puce	Paramètre d'entrée	Dimensions de la puce
H_M, V_M	Espacements horizontal et vertical des blocs dans une puce	Géométrie de la puce	Paramètre d'entrée	Dimensions de la puce
σ_m	Variance de la loi N (0,σm²)	Variations horizontale et verticale des distances inter-bloc	H_M, V_M	-
P_{de}, P_{do}, P_{di}	Pourcentages des différentes formes de spots	Géométrie de la puce	Paramètre d'entrée	-
M_h, M_v	Marges verticales et horizontales	Géométrie de la puce	Paramètre d'entrée	Position de la puce dans l'image

Tableau 2-1 : paramètres géométriques du modèle

2.2. Modélisation du signal

Une quantification fiable des niveaux d'expression des gènes passe par la maximisation du rapport signal sur bruit (RSB) dans les images. Pour modéliser fidèlement le système d'acquisition utilisé dans la formation des images de puces, il est indispensable d'identifier précisément les mécanismes susceptibles de baisser ce rapport RSB, et donc d'identifier les sources potentielles de bruit.

2.2.1. Différentes sources de bruit

Trois sources principales de bruit peuvent être identifiées dans des images de puces : le bruit électronique, généré par les appareils et les fuites de courant ; le bruit expérimental, dû à la fois au substrat (diffusion, réfraction) et à l'expérience elle-même (lavage inefficace du substrat, contamination par des poussières ou des impuretés causant des fluorescences non désirées, hybridations non spécifiques) ; et enfin un bruit causé par l'étroite dépendance entre le signal et le nombre de photons détecté (la variabilité du nombre de photons croît comme le nombre de photons incidents)
Dror [DROR01] propose une étude complète des différents bruits rencontrés dans des images de puce, selon leur nature multiplicative (intensité du bruit proportionnelle à l'intensité du signal) ou additive

(intensité du bruit indépendante), et selon la nature de leur action (globale ou locale au dépôt). Le Tableau 2-2 résume l'ensemble de ses conclusions.

	Multiplicatif	Additif
Action globale	• Variation des durées d'hybridation • Variation des concentrations de matériels biologiques • Lavage inefficace du substrat	• Variation de luminosité durant la lecture de la puce
Action locale	• Inhomogénéité des préparations biologiques • Diffusion, réfraction • Variations des intensités laser	• Contaminations • Hybridations croisées

Tableau 2-2 : classification des bruits d'une image de puce à ADN (source : [DROR01])

Le simulateur se propose de modéliser l'ensemble de ces phénomènes, ainsi que le "vrai" signal d'expression des gènes, en gérant à la fois les paramètres physiques impliqués dans la formation de l'image, et les paramètres biologiques inhérents aux préparations des sondes, des cibles et du substrat.

2.2.2. Le signal du fond

Les scanners utilisent pour la plupart un laser permettant de stimuler la fluorescence des molécules marquant les sondes sur le substrat. Connaissant la longueur d'onde caractéristique des lasers utilisés (de l'ordre de 600 nm), le substrat satisfait toujours le critère de Rayleigh, et la réponse du scanner est du type speckle, modélisant la réflexion d'une onde cohérente sur une surface plane. La modélisation du signal du fond adoptée ici est celle des diffuseurs discrets [GOODMAN76], qui stipule que le substrat est constitué d'un milieu homogène dans lequel sont distribués des diffuseurs discrets et identiques. Selon la densité des diffuseurs dans la cellule de résolution et leur loi de distribution (aléatoire ou non), le comportement statistique de l'image de fond varie.

- Modèle de type Rayleigh

Le modèle de Rayleigh suppose que la distribution spatiale des diffuseurs dans le milieu est aléatoire et non corrélée. Le signal rétrodiffusé peut être alors modélisé comme une somme de toutes les réponses individuelles des diffuseurs dans la cellule de résolution à un instant donné. Autrement dit, le signal résulte d'une somme vectorielle complexe de composantes aléatoires.
La réponse de chaque diffuseur est un vecteur, appelé phaseur aléatoire, d'amplitude a_i et de phase φ_i (due à sa localisation spatiale aléatoire) et la réponse totale R d'un tel milieu exploré par le laser est :

$$R = A.exp(j\Theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} a_i e^{j\varphi_i}$$

où n est le nombre de diffuseurs dans une cellule de résolution donnée.

En supposant que les φ_i suivent une loi uniforme sur $[-\pi,\pi]$, et que les a_i sont indépendantes, les parties réelle et imaginaire de la réponse R sont deux variables aléatoires décorrélées de moyenne nulle et de variance donnée par :

$$E(R_r^2) = E(R_i^2) = \sigma^2.$$

Si les amplitudes a_i sont supposées constantes, la modélisation prend en compte l'homogénéité du substrat ; dans le cas contraire, la variabilité modélise par exemple un mauvais processus de lavage de la lame. Si de plus le nombre n de diffuseurs dans une cellule de résolution devient important, les distributions des parties réelle et imaginaire de R suivent une loi normale $N(0, \sigma^2)$ en vertu du théorème centrale limite.
Le signal effectivement lu étant le module du signal complexe R, la densité de probabilité de l'amplitude détectée A peut être obtenue après changement de variable des coordonnées cartésiennes en coordonnées polaires :

$$P(A)=\frac{A}{\sigma^2}e^{-\left(\frac{A}{2\sigma^2}\right)}, A \geq 0,$$

qui est une distribution de Rayleigh dont une des propriétés remarquables est d'avoir un rapport signal sur bruit constant de 1.91 pour un grand nombre de diffuseurs considéré dans une cellule de résolution donnée.

Ce modèle suppose la présence d'un grand nombre de diffuseurs dans une cellule de résolution, et le signal détecté est alors qualifié de speckle pleinement développé. Quand le nombre de diffuseurs est restreint (typiquement inférieur à 20), le rapport signal sur bruit a une valeur inférieure à 1.91, ce qui rend le modèle de Rayleigh non général pour la description du signal à modéliser.

- Modèle de type Rice

Le modèle de Rayleigh échoue lors de la présence d'une composante cohérente due à la présence d'une structure régulière de diffuseurs sur la lame. Le modèle de phaseurs aléatoires peut alors être modifié en :

$$R=S+A.exp(j\Theta),$$

où S représente la partie cohérente. La densité de probabilité du signal détecté est alors de la forme

$$P(A)=\frac{A}{\sigma^2}e^{-\left(\frac{A^2+S^2}{2\sigma^2}\right)}I_0\left(\frac{SA}{\sigma^2}\right), A \geq 0$$

où I_0 est la fonction de Bessel de première espèce et d'ordre 0. Cette densité est une distribution de Rice, dont le rapport signal sur bruit augmente linéairement avec le rapport S/σ.

Ces deux modèles sont utilisés dans la suite pour simuler le signal de fond d'une image de puce à ADN, en relation avec la physique d'acquisition. Ils autorisent de plus la simulation de contaminations (poussières, cheveux…) : ainsi un pixel de poussière peut être modélisé par un seul réflecteur R_0 d'amplitude importante a_0 et de phase φ_0 nulle.

2.2.3. Le signal des spots

Les fluorochromes présents sur un dépôt absorbent les photons d'excitation générés par le laser, et émettent en réponse des photons de fluorescence, cette fluorescence étant une fonction croissante du processus d'hybridation. On considère de plus que le nombre de photons réémis, et donc l'intensité lue correspond de façon linéaire à la quantité de fluorochrome présent, même si les scanners ne sont pas des systèmes parfaitement linéaires sur l'ensemble de la gamme lue. Les photons de fluorescence peuvent se diriger dans n'importe quelle direction, et seule une fraction d'entre eux est collectée par une lentille. Des photomultiplicateurs convertissent alors ces derniers en un courant électrique, finalement numérisé par une convertisseur analogique/numérique.

L'émission de ces photons et leur interaction avec leur environnement sont des processus hautement aléatoires. De nombreux auteurs ont proposé des analyses statistiques du signal des spots [HOYLE02] [KERR00] [RUDEMO02], et la modélisation retenue ici pour le signal des sondes dans un spot (et donc des fluorochromes) utilise une loi de Poisson. Puisque toutes les émissions sont indépendantes pour un spot donné, et suivent la même loi statistique, le signal global rétro-diffusé d'un spot suit une loi gaussienne, dont la moyenne dépend du nombre de molécules hybridées, de la puissance du laser et de la sensibilité du détecteur. La simulation de la distribution des intensités dans un spot suit donc ici une loi normale $N(m_s,\sigma_s^2)$. Supposant de plus une corrélation forte entre les spots déposés par la même aiguille [SCHUCHHAR00], et sachant qu'une aiguille donnée dépose tous les spots d'un même bloc, les paramètres

de la loi normale ne doivent dépendre que du bloc simulé. La moyenne de cette loi, définissant l'expression du gène considéré, aura une valeur soit aléatoire, soit paramétrique (permettant par exemple d'introduire des dépendances entre gènes pour l'étude de l'étape de classification), soit réelle (récupérée à partir d'expériences *in vivo*).
Lorsqu'il s'agit de comparer deux populations de cellules, hypothèse est enfin faite que les fluorescences détectées dans les deux canaux de fluorescence (*e.g.* vert et rouge) sont identiques à quelques variations près : la géométrie des spots est la même, et le niveau d'expression rouge (resp. vert) suit une loi normale $N(m_r, \sigma_r^2)$ (resp. $N(m_g, \sigma_g^2)$), m_r et m_g (resp. σ_g et σ_b) étant lies par une relation de la forme :

$m_r = m_g + \delta_m$, δ_m suivant une loi $N(0, \alpha m_g)$
$\sigma_r = \sigma_g + \delta_\sigma$, δ_σ suivant une loi $N(0, \alpha\sigma_g)$,

α étant un pourcentage.

L'ensemble de cette modélisation génère pour chaque spot géométrique une activité simulant le niveau d'expression du gène correspondant. Cependant, en raison d'un manque éventuel d'hybridation spécifique et/ou d'une anomalie de détection, certains spots sont très sous exprimés sur des images réelles de puces à ADN. A l'inverse, certains autres peuvent être sur exprimés et saturer la dynamique de l'image. Le modèle intègre ces cas d'une part par l'intermédiaire d'un pourcentage P_{ue} de spots sous exprimés, et d'autre part en imposant à un certain nombre de spots (pourcentage P_{oe}) une activité moyenne m_r (ou m_g) proche du niveau de saturation. La localisation de ces deux types de spots est corrélée avec des défauts de dépôt par les aiguilles du robot.

2.2.4. Non uniformité de la puce

Les systèmes d'acquisition scanner utilisent typiquement un système confocal avec deux chemins optiques, l'un excitateur (du laser vers la puce) et l'autre émetteur (sens inverse). Le chemin excitateur a une ouverture relativement faible, tandis que l'émetteur nécessite de maximiser la collecte des photons et doit donc avoir la plus large ouverture possible. Le compromis passe par alors une profondeur de champ du système optique faible, qui peut engendrer un défaut de focalisation du le support. A cet effet est ajouté une planéité imparfaite du support de la puce [HANDRAN01], le tout produisant une hétérogénéité du signal indépendante des produits biologiques déposés. Cette variabilité est simulée en multipliant l'image d'activation des spots par une carte de variation basses fréquences des niveaux de gris, d'amplitude maximale *M*, et dont la forme peut être plane, exponentielle, parabolique, en gradient (*X* ou *Y*) ou de forme aléatoire (Figure 2-4).

Parabolique

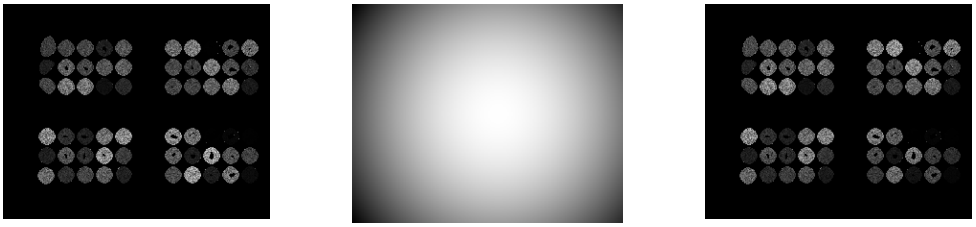


Image originale Exemples de cartes de variation Exemple de résultat

Figure 2-4 : exemples de cartes de variation basses fréquences des niveaux dans l'image.

Le Tableau 2-3 résume finalement l'ensemble des paramètres de signal du modèle.

Nom	Description	Fonction	Lié à
n	Nombre de diffuseurs	Simulation du bruit de fond	Paramètre d'entrée
P_{du}	Pourcentage de poussière	Simulation de la présence de poussière	Paramètre d'entrée
A	Intensité moyenne du diffuseur	Simulation du bruit de fond	Paramètre d'entrée
m_r	Niveau moyen d'expression rouge	Simulation de l'activité dans le rouge	Paramètre d'entrée
σ_r	Variance d'expression rouge	Simulation de l'activité dans le rouge	Paramètre d'entrée
α	Pourcentage	Dépendance intra-bloc rouge/vert	Paramètre d'entrée
P_{ue}	Pourcentage de spots sous exprimés	Modélisation de la disparition de spots	Paramètre d'entrée
P_{oe}	Pourcentage de spots sur exprimés	Modélisation de la saturation des spots	Paramètre d'entrée
M	Amplitude de variation du signal	Simulation de l'hétérogénéité du signal	Paramètre d'entrée
T	Type de carte basses fréquences	Description de l'hétérogénéité du signal	Paramètre d'entrée

Tableau 2-3 : paramètres de signal du modèle

3. Résultats et discussion

3.1. Validation

3.1.1. Modélisation géométrique

Nous avons établi les différents modèles géométriques en fonction de discussions avec des biologistes (laboratoire de biologie des protistes, équipe Génomique Intégrée des Interactions Microbiennes, Pierre Peyret, UMR 6023 CNRS). Ces modèles semblent reprendre l'essentiel des artefacts de formation des spots rencontrés sur de vraies images de puce à ADN. Nous n'avons sciemment pas reproduit des artefacts tels que les queues de comètes (Figure 2-5), dus à un retrait du support avant séchage complet ou à une immersion trop lente du support dans la solution fixante, les images générées alors étant totalement inexploitables.

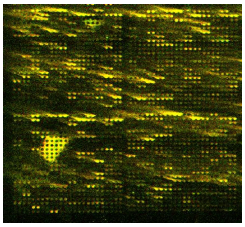
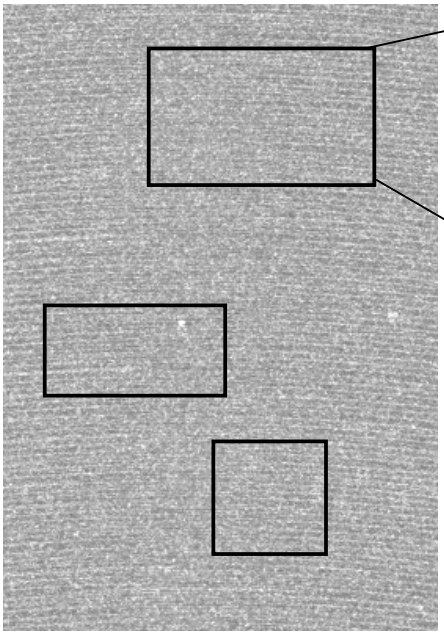
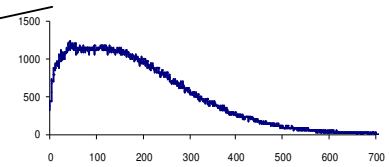


Figure 2-5 : queues de comètes

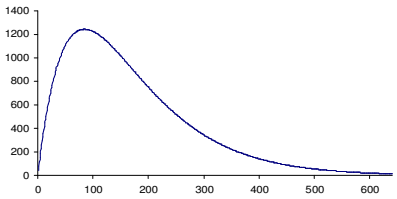
3.1.2. Modélisation du signal

De nombreuses expériences ont été réalisées à l'aide d'un matériel disponible à l'Université (système Affymetrix 418), essentiellement pour vérifier la modélisation statistique du signal du fond sur ce type de puces. Des supports de verre sans dépôt biologique significatif ont été imagés en utilisant le même protocole d'acquisition, et des régions d'intérêt ont été extraites des images résultantes pour mesurer la statistique du fond (Figure 2-6).





Histogramme d'une région d'intérêt
Paramètres : moyenne : 193.37, médiane : 167.5
Variance: 22286.4, coefficient de variation : 0.77
Kurtosis : 486.06



Distribution de la loi de Rayleigh

Régions d'intérêt dans le fond

Figure 2-6 : distribution des intensités dans le fond

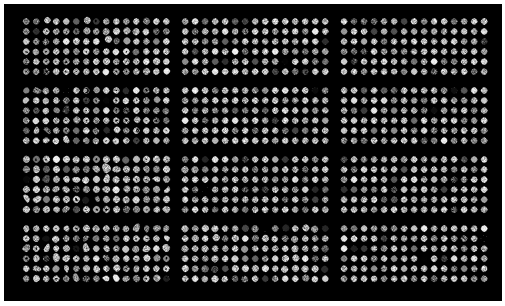
L'histogramme des intensités dans les régions d'intérêt a plus particulièrement été comparé à une distribution de Rayleigh. Pour ce faire, un test du χ^2 a été utilisé : la gamme de variations des niveaux de gris dans la région d'intérêt a été répartie dans k classes distinctes, et la quantité

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

a été calculée, O_i étant la fréquence observée pour la classe i, E_i la fréquence théorique observée. La loi de Rayleigh étant monoparamétrique, l'hypothèse selon laquelle la loi statistique du fond observé est une loi de Rayleigh doit dont être rejetée si $\chi^2 > \chi^2_{\alpha,k-2}$, α étant la tolérance et $\chi^2_{\alpha,k-2}$ une valeur tabulée.
Pour toutes les expériences effectuées, et ce tant que les cardinalités des classes étaient suffisantes (supérieures à 5 individus, critère indispensable pour l'application du test [SAPORTA91]), l'hypothèse n'a pu être rejetée.

3.2. Quelques résultats

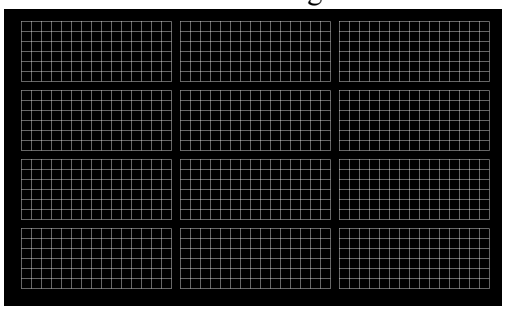
Le simulateur a été réalisé en langage C et des versions sous environnement UNIX et Windows sont disponibles. Il est paramétré *via* un simple fichier texte, collectant l'ensemble des paramètres géométriques et de signal. Le calcul d'une image définie par les paramètres géométriques (R_s=5µm, S_H=15, S_V=6, H_B=15µm, H_V=25µm, B_H=4, B_V=3, H_M=50µm, V_M=50µm, P_{de}=P_{do}=P_{di}=0.2) et de signal (n=40, P_{du}=10^{-5}, A=10, m_r =3000, σ_r =20, α=0.15, P_{ue}=0.15, P_{oe}=0.15, T =parabolique, M=1) est effectué en 4 secondes sur un Pentium IV 2.4GHz disposant de 512Mo de RAM. La Figure 2-7 présente cet exemple complet et toutes les sorties proposées par le simulateur.



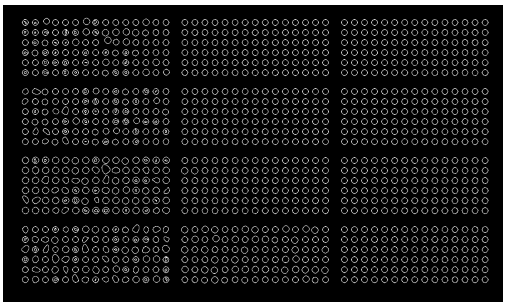
a - Image

R_s = 5 µm	n = 40
S_H = 15	$P_{du} = 10^{-5}$
S_V = 6	A = 10
H_B = 15 µm	m_r = 3000
H_V =25 µm	σ_r = 20
B_H = 4	α = 0.15
B_V = 3	P_{ue} = 0.15
H_M = 50 µm	P_{oe} = 0
V_M = 50 µm	T = parabolique
$P_{de} = P_{do} = P_{di} = 0.2$	M = 1
Paramètres géométriques	Paramètres de signal

b - Paramètres



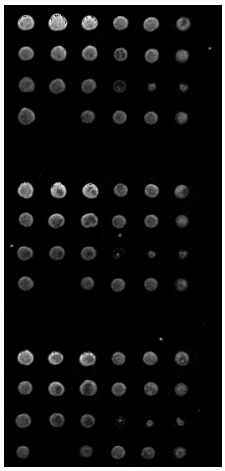
c – Plan de dépôt théorique



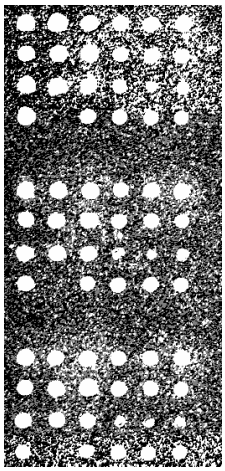
d - Géométrie de la puce

Figure 2-7 : premier exemple de simulation

La Figure 2-8 présente quant à elle une illustration de la modélisation du bruit de fond. La dynamique de l'image (a), comportant par ailleurs des formes de spots variées, a été étalée par un algorithme d'égalisation d'histogramme, pour faire ressortir les caractéristiques du bruit et l'hétérogénéité du signal sur le support.



a – Simulation



b – Egalisation d'histogramme

Figure 2-8 : second exemple de simulation

3.3. Applications

Nous avons relié l'ensemble des paramètres du simulateur soit à des considérations géométriques, soit à des données relatives au signal, dans le but de rendre ce modèle utilisable par le plus grand nombre (informaticiens ou non), la bioinformatique touchant un public très varié. Le modèle est de plus suffisamment flexible pour intégrer d'autres paramètres spécifiques (*e.g.* correspondant à d'autres types de substrat comme le nylon).

3.3.1. Validation d'outils existants

Nous avons tout d'abord pensé ce simulateur comme un outil de validation de méthodes couramment utilisées en analyse des puces existantes. Il est envisageable qu'une fois validé, ce modèle puisse être utilisé au laboratoire comme un outil :

- de calibration de méthodes mises à la disposition des biologistes pour l'analyse des puces à ADN, en proposant en particulier des domaines de conformité pour les différents outils utilisés, et ce en fonction de diverses conditions d'acquisition, de préparation biologique et de pré-traitement, le but étant par exemple de proposer des logiciels "à façon" en fonction des spécificités de chacun
- de validation de toute nouvelle méthode d'analyse, à quelque niveau que ce soit (segmentation des spots, classification,..) en permettant de comparer les résultats de cette méthode à la réalité donnée par les paramètres du simulateur.

A notre sens, ces deux points sont fondamentaux, puisque l'analyse informatique de la puce a toujours un impact considérable sur l'interprétation biologique des données. Le reste de ce manuscrit se consacre au développement de ces deux aspects.

Outre son aspect flexible (car paramétrique), cette modélisation doit plus généralement permettre d'effectuer des simulations de puces à un coût nul, contrairement à une fabrication de puce réelle qui représente un coût non négligeable (il faut actuellement compter entre 20 et 30 euros pour la synthèse d'un oligonucléotide de 50 bases).

3.3.2. Plateforme WEB

Nous souhaitions également proposer ce simulateur d'images de puces à ADN à la communauté bioinformatique, pour que cet outil puisse effectivement être validé par le plus grand nombre. Une version WEB du simulateur est donc également accessible à l'adresse Internet http://www.isima.fr/bioinfo/Logiciels/Simulateur/. L'intégration a été réalisée à l'aide de pages dynamiques utilisant les technologies HTML, PHP et JavaScript. Trois formulaires permettent à l'utilisateur de donner des informations personnelles (pour les statistiques d'utilisation du site), puis de spécifier les paramètres de l'image souhaitée (Figure 2-9). La soumission de l'ensemble de ces formulaires lance la simulation et génère les images de puce résultantes (dans les deux canaux rouge et vert), un courriel étant automatiquement envoyé à l'utilisateur pour lui proposer le téléchargement des données. Les images sont en format TIFF 16 bits sans compression, permettant ainsi leur utilisation sur tous les logiciels classiques d'analyse. La description du plan théorique de dépôt (au format CSV), ainsi qu'un fichier des « vraies » activations des gènes simulées (pour la référence) sont également proposés en téléchargement.

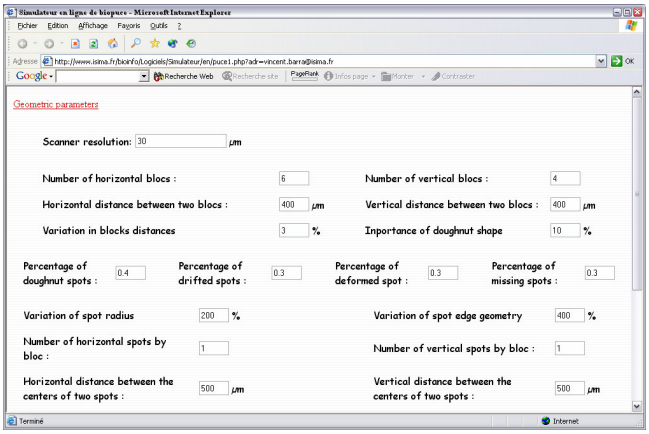


Figure 2-9 : capture d'écran du simulateur en ligne

Conclusion

L'interprétation biologique des expériences issues de puces à ADN étant intrinsèquement liée à l'analyse des images mises en jeu, il est essentiel de pouvoir valider l'ensemble du processus de traitement pour fiabiliser les résultats. Ce chapitre a proposé un simulateur d'images de biopuces transcriptome, combinant une modélisation des facteurs de variations affectant la géométrie et le signal des images acquises. Les deux versions du simulateur proposé, autonome et client-serveur en ligne, donnent un accès possible à la validation des différentes étapes de traitement des images de biopuces, et font l'objet de deux publications récentes [BARRA03][BARRA04-B].

Les chapitres suivants se proposent d'étudier plus précisément les algorithmes utilisés pour l'analyse des images de puce à ADN, et de tester leurs performances à l'aide de ce simulateur.

Chapitre 2 - Analyse d'images de puces à ADN

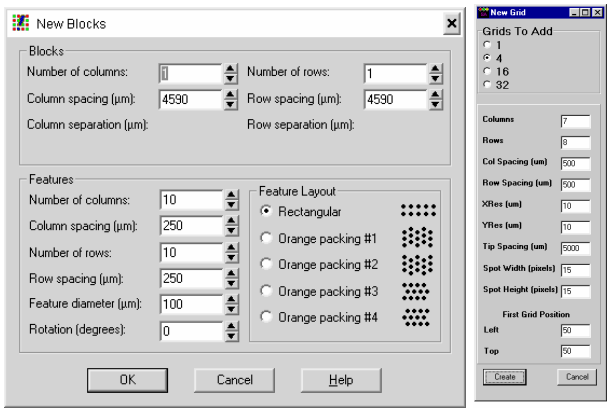
On dessine toujours les éléphants plus petits que nature,
mais les puces sont toujours plus grandes.
Jonathan Swift

Introduction
Une fois les données effectivement acquises, l'étape de traitement d'images consiste à extraire l'information biologique pertinente, qui sera exploitée *a posteriori* et de laquelle émergeront des conclusions quant aux réseaux de régulation et aux rôles des gènes étudiés, ou encore concernant les relations entre ces mêmes gènes. L'analyse des images de biopuces est généralement effectuée en quatre étapes : une première phase de repérage de la géométrie de la puce (l'adressage), une seconde de segmentation des spots, une troisième de quantification des niveaux d'activité des gènes présents sur la puce, et une dernière étape d'extraction de connaissances biologiques à partir de ces données quantitatives. Si le chapitre 3 s'intéresse plus particulièrement à ce dernier point, le présent chapitre développe l'ensemble des trois premières étapes du processus d'analyse.
Les méthodes classiques sont tout d'abord rappelées puis évaluées à l'aide du simulateur présenté dans le chapitre précédent. Pour chacun des facteurs de variations recensé, les algorithmes mis en œuvre dans les logiciels dédiés sont testés, et leurs limites sont démontrées. Un nouvel algorithme réalisant simultanément l'ensemble des trois étapes du processus d'analyse d'images est ensuite proposé, fondé sur une division itérative de l'image en triangles élémentaires, en fonction de critères d'homogénéité du signal. Cette méthode est validée sur des données de synthèse proposées par le simulateur, et démontre une bien meilleure robustesse aux différentes variabilités observées sur de réelles images.

1. Adressage

La structure géométrique de base d'une image de puce à ADN est déterminée par le robot, et est donc *a priori* connue. Cependant, comme précisé dans le chapitre 1, certains paramètres doivent être estimés, parmi lesquels les espacements horizontaux et verticaux entre les blocs, les mouvements affines des blocs et des spots, les espacements horizontaux et verticaux entre les spots d'un même bloc ou encore la position des grilles par rapport aux bords de l'image.
Les variations de ces variables font que, le plus souvent, le plan de dépôt théorique ne correspond pas exactement à la configuration géométrique observée sur l'image de puce. D'où la nécessité de procéder à une étape d'adressage, dans laquelle une grille régulière va être calculée sur l'image, et où chaque case ne contiendra qu'un et un seul spot. Outre son intérêt pour l'étape suivante de segmentation, cette phase donne un repère géométrique qui, dans le cas de l'acquisition en double marquage, permet un recalage précis d'une image sur l'autre.
Bon nombre de méthodes automatiques ou semi-automatiques résolvant ce problème ont été publiées, utilisant des techniques empruntées à l'analyse d'images. Elles incluent l'analyse des profils horizontaux et verticaux de images par transformée de Fourier [JOUENNE01], par fonctions de scores [JAIN02] ou par filtrage [LAWS03], l'utilisation de méthodes morphologiques [SIDDIQUI03], statistiques [JOUENNE01] [LIEW03] ou encore l'analyse de l'image par classifieurs [JUNG02] ou par champs de Markov [KAZTER03]. Des logiciels dédiés à l'analyse des images de puces à ADN implémentent certaines d'entre elles (*e.g.* DigitalGenome-Analyzer DG, MolecularWare ; SPOT, CSIRO Mathematical and Information Sciences), mais globalement, ces méthodes ne se révèlent pas assez robustes dans le cas d'images de mauvaise qualité, et l'intervention d'un opérateur qui réajuste la grille est alors indispensable. Dans la plupart des cas, d'ailleurs, la grille théorique est fournie lors du plan de dépôt, par exemple sous la forme d'un fichier CSV (Comma-Separated Value), et l'utilisateur a la possibilité de déplacer et déformer cette dernière (Jaguar, Affymetrix ou GenePix, Axon) Il est également possible de définir la grille par ses

paramètres géométriques à l'aide de boîtes de dialogue adaptées (Figure 2-10). L'ajustement manuel des cases de la grille peut bien sûr devenir vite fastidieux et très coûteux en temps, et sujet à de grandes variabilités inter utilisateurs.



GenePix ScanAlyze

Figure 2-10 : exemples de boîtes de dialogue pour la définition de grilles

D'une manière générale, le positionnement de la grille doit être simple, rapide à réaliser, et aussi juste que possible. En effet, de la précision de cette étape dépend l'efficacité des mesures ultérieures, en particulier concernant la segmentation des spots.

2. Segmentation

L'étape de segmentation des images de puces à ADN est la phase cruciale lors de laquelle les données à analyser, et reliées à des propriétés biologiques des échantillons étudiés, sont effectivement générées.
Le principe de cette étape est d'isoler dans chacune des cases définies à l'étape d'adressage les pixels des spots des pixels du fond, *i.e.* réaliser une segmentation en deux classes. Yang *et al.* [YANG00] ont classé quelques méthodes de segmentation en quatre groupes, en fonction de la géométrie des spots qu'elles se proposent de segmenter : les méthodes à cercles fixes ou adaptatifs, les méthodes à géométrie variable, et les techniques fondées sur l'histogramme. Nous proposons ici une autre classification, en fonction de la nature de la segmentation réalisée.

2.1. Segmentation spatiale

Ce type de segmentation correspond à la classe de méthodes « cercle fixe » proposée dans [YANG00]. Sans tenir compte de l'information du signal dans l'image, il s'agit ici d'ajuster un cercle à diamètre constant sur tous les spots de l'image. Proposé pour la première fois dans le logiciel ScanAlyze (Michael Eisen, Université de Berkeley, Californie), ce type d'algorithme est depuis proposé en option sur de nombreux logiciels d'analyse. Facile à implémenter, rapide à utiliser, ce type de technique ne fonctionne cependant que lorsque tous les spots sont parfaitement définis et rigoureusement identiques, ce qui n'est en pratique jamais le cas.

2.2. Segmentation en intensité

Dans cette classe de méthodes, seule l'information en signal portée par l'image est utilisée pour segmenter les spots. L'hypothèse sous-jacente est que les niveaux de gris des pixels des spots sont statistiquement différents des pixels du fond. Les méthodes les plus rencontrées ici sont du type analyse d'histogramme (QuantArray, GSI Lumonics et ImaGene, Biodiscovery) pour lesquelles un masque générique, de taille supérieure à la taille d'un spot, est appliqué sur l'image, et dans lequel un histogramme des niveaux de gris est calculé pour déterminer des seuils d'appartenance au fond ou au signal. Si cette méthode est facile à réaliser et rapide d'utilisation, elle est peu robuste et dépend fortement du masque arbitrairement défini. D'autres techniques ont été mises en place qui comblent partiellement ces lacunes, les principales utilisant des techniques de croissance de régions (Spot), de détection de

contours (utilisation du Laplacien dans le logiciel Dapple [KIM01], de méthodes de géométrie différentielle [DEMONGEOT02]), des méthodes de classification de signal ([BOZINOV02] [ERGÜT03] [NAGARAJAN03]), de morphologie mathématique ([ANGULO03] [SIDDIQUI03]), de modélisation statistique ([BRÄNDLE00], [LIEW03]) ou encore l'algorithme des watershed ([LOTUFO03]).

2.3. Méthodes mixtes

L'idéal est de construire des algorithmes exploitant à la fois l'information spatiale régulière imposée par le plan de dépôt et la géométrie des spots, et le signal de l'image. Le principe général repose tout d'abord sur une localisation grossière de la position des spots à l'aide de cercles (ou d'ellipses [LAWS03]), puis sur une amélioration de la segmentation des bords des spots en utilisant des considérations sur le signal (suppression des outliers à l'intérieur et à l'extérieur des cercles).

Les méthodes de cette catégorie, les plus rencontrées dans la littérature et sur les logiciels dédiés utilisent une géométrie en cercle adaptatif (GenePix), un test de Mann et Whitney ([CHEN97], QuantArray et DeArray, Scanalytics) ou encore une croissance de régions contrôlée (*e.g.* détermination de la position des germes contrôlée par la géométrie dans ImaGene).

Globalement, les algorithmes aujourd'hui proposés dans les logiciels dédiés sont souvent des boîtes noires peu évolutives, très sensibles aux divers artefacts et doivent donc souvent être affinés à la main pour une quantification satisfaisante. Ils ne fonctionnent souvent bien qu'avec certaines combinaisons de robots/scanners, ou seulement pour certaines conditions expérimentales, et il est très délicat de proposer un algorithme « générique » traitant tous les cas.

3. Quantification

La dernière étape avant l'exploitation effective des données est la phase de quantification, durant laquelle chaque spot maintenant identifié se voit résumer par une valeur numérique (ou un ratio de valeurs numériques dans le cas d'une expérience à double marquage). Durant cette phase, une quantification du signal du fond est également envisagée, pour une étape ultérieure de correction des expressions avant interprétation biologique.

3.1. Intensité des spots

Sous certaines conditions, il est possible de considérer que l'intensité d'un spot est directement proportionnelle au niveau d'expression du gène considéré :

- la quantité de sonde déposée sur chaque spot est constante ;
- la contamination des spots est inexistante ;
- le scanner utilisé est un système parfaitement linéaire sur l'ensemble de la gamme lue.

Généralement, au moins une de ces conditions n'est pas respectée, mais il est d'usage de considérer que la valeur représentative que l'on souhaite extraire des spots représente de manière synthétique la quantité de fluorochrome présent et donc le niveau d'hybridation du gène correspondant.

Le problème posé ici concerne alors la recherche d'une valeur représentative à retenir pour chaque spot. Quatre indices sont généralement retenus :

- la médiane, qui n'est influencée par les pixels d'intensité extrême que par leur nombre, et non leur intensité. Cet indice est donc peu sensible aux outliers ;
- la moyenne m, dont le principal inconvénient est sa dépendance aux intensités extrêmes (pixels du fond mal segmentés, pixels de saturation) ;
- la somme des intensités des pixels du spot, très sensible aux problèmes de contamination et de mauvaise segmentation (outliers) ;
- et enfin le mode du signal qui est une valeur représentative du pic de l'histogramme des niveaux de gris du spot.

Si aucun indice n'émerge dans la littérature comme étant supérieur aux autres, la plupart des logiciels dédiés proposent toute une gamme d'indices statistiques à partir de ces quatre valeurs de base (écart type e des niveaux de gris dans chaque spot et dans le fond, éventuellement sur les deux canaux, nombre de

pixels hors de l'intervalle $m \pm e$, moyenne ou médiane des ratios, somme des moyennes ou des médianes, ratio des corrélations entre les deux canaux…)

3.2. Intensité du fond

Dans le but d'estimer une « vraie valeur » d'expression des gènes, il est nécessaire de réduire l'effet des fluorescences non spécifiques, et une estimation de l'intensité du fond doit donc être effectuée. Hypothèse est faite ici que le signal du fond se manifeste sous la forme d'un bruit additif.
La correction du fond peut être envisagée sous quatre angles :

- en considérant pour chaque spot le signal du fond voisin : une valeur représentative du fond (médiane, moyenne) est calculée pour chaque spot à partir de ses pixels voisins, puis soustraite des intensités du spot (Jaguar). Statistiquement, cette méthode fournit un estimateur peu robuste du fond, la taille des échantillons permettant de calculer la valeur représentative étant faible ;
- en utilisant un voisinage des spots voisins : ici plusieurs techniques d'analyse sont utilisées, comme des noyaux de convolution, des filtres de rang ou adaptatifs (*e.g.* ouverture morphologique dans le logiciel Spot). L'estimation de l'intensité du fond est alors plus robuste ;
- en traitant le fond de manière globale dans toute l'image, le bruit de fond étant alors estimé à l'aide d'une fonction paramétrique [KOOPERBERG02] ;
- En considérant qu'aucune correction ne doit être faite !

Selon Yang *et al.* [YANG00], le choix de la méthode d'estimation du signal du fond a autant d'importance que la précision de la segmentation des spots pour déterminer un niveau d'expression fiable des gènes.

3.3. Qualité des mesures

En plus des intensités des spots et du fond, il est également souhaitable de disposer de statistiques décrivant la qualité des mesures collectées [WANG01]. Les logiciels dédiés proposent cette option sous la forme de rapports signal/bruit ou de mesures sur les spots (surface pour une validité statistique de la médiane par exemple), ainsi que des tests de rejet de la qualité des spots. A titre d'illustration, Dapple propose deux indices, un b-score mesurant la fraction des intensités du fond inférieures à l'intensité médiane des spots, et un p-score évaluant la dérive moyenne de la position du spot par rapport au plan de dépôt rectangulaire initial.

Nous proposons dans la suite une méthode initiale de segmentation, donnant directement accès à une recherche automatique de grille et à la quantification des activités des spots de l'image [BARRA04-C].

4. Méthode proposée

4.1. Principe

Nous proposons ici d'associer le problème de segmentation à la recherche d'une partition de l'image en triangles élémentaires, obtenus par un processus de triangulation automatique. Le processus est de type division/fusion et se propose, à partir d'une triangulation initiale, de converger en itérations vers une partition où chaque maille élémentaire représente une approximation d'une partie de l'image selon des critères d'ordre statistique.

4.2. Triangulation automatique

L'algorithme de segmentation proposé nécessite tout d'abord de trianguler automatiquement le domaine de l'image. Nous souhaitons disposer d'une triangulation dans laquelle les triangles ont une forme relativement homogène, ceci notamment dans le but de faciliter les étapes ultérieures de division/fusion, et nous proposons donc ici de calculer une triangulation de Delaunay associée à un ensemble de sommets répartis sur l'image.

4.2.1. Triangulation de Delaunay

Soit $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ un ensemble de points du plan euclidien E. Sous l'hypothèse que trois points quelconques de $\mathcal{P}$ ne sont jamais colinéaires et que 4 points de $\mathcal{P}$ ne sont jamais cocirculaires, la région :

$R_i = \{x \in E^2 / d(x,P_i) < d(x,P_j)\ ,\ \forall j \neq i\}$,

où $d(x,P_j)$ désigne la distance entre x et P_j, est appelée cellule de Voronoï associée au germe P_i. Le pavage du plan formé par l'ensemble des R_i est le diagramme de Voronoï associé à $\mathcal{P}$. La construction $DT(\mathcal{P})$ qui consiste à rejoindre les points P_i des cellules de Voronoï voisines est nommée triangulation de Delaunay. Cette triangulation est telle qu'aucun triangle $P_iP_jP_k$ de $DT(\mathcal{P})$ ne contient un autre point de $\mathcal{P}$ à l'intérieur de son cercle circonscrit.
Par construction, la triangulation de Delaunay assure que parmi toutes les triangulations de E, DT($\mathcal{P}$) est celle qu maximise l'angle minimum de tous les triangles, assurant une homogénéité dans la forme de ces derniers (et évite en particulier le cas de triangles quasi plats, gênants pour la suite de l'algorithme).
De nombreux algorithmes permettent de calculer la triangulation de Delaunay d'un nuage de points $\mathcal{P}$. On trouve en particulier un algorithme récursif de type "diviser pour régner" extrêmement performant. Néanmoins, les méthodes récursives ne sont utiles que quand le nuage de points est connu à l'avance. Dans le cadre de l'approche division/fusion retenue, $\mathcal{P}$ est amené à évoluer au cours du temps et nous orientons donc la construction de $DT(\mathcal{P})$ vers une méthode incrémentale en choisissant l'approche proposée par Watson [WATSON81].

4.2.2. Algorithme de Watson

Dans un premier temps, les points de $\mathcal{P}$ sont englobés dans un triangle ABC, permettant d'initialiser la méthode ($DT(\mathcal{P})=\{(ABC)\}$). Les sommets P_i sont ensuite ajoutés successivement :

- le triangle T contenant P_i est recherché et trois nouveaux triangles sont formés en reliant P_i aux trois côtés de T, qui est supprimé de $DT(\mathcal{P})$. Le gain net en triangles est donc de deux unités ;
- les triangles adjacents à T sont placés dans une pile. Pour chaque triangle T_j de cette pile, on vérifie si P_i appartient à son cercle circonscrit. Dans ce cas, T_j et l'un des nouveaux triangles T_k forme un quadrilatère avec une diagonale placée dans le mauvais sens. T_j et T_k sont alors remplacés par deux nouveaux triangles obtenus en permutant la diagonale du quadrilatère (Figure 2-11). Lorsque la permutation est effectuée, les triangles adjacents aux deux nouveaux triangles et qui n'ont pas P_i comme sommet sont ajoutés à la pile, le processus itérant jusqu'à ce que la pile des triangles soit vide.

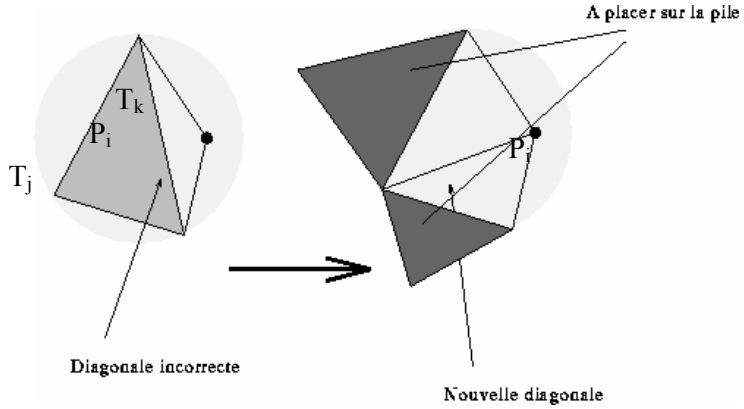


Figure 2-11 : construction itérative d'une triangulation de Delaunay

Un final, $2n+1$ triangles sont créés, incluant ceux contenant les sommets du triangle ABC. La triangulation de Delaunay $DT(\mathcal{P})$ est obtenue en détruisant tous les triangles qui contiennent un ou plusieurs sommets A, B ou C.
En utilisant des structures de données adaptées au stockage de $DT(\mathcal{P})$ et aux différentes étapes de l'algorithme de Watson (notamment à la recherche du triangle contenant un point de $\mathcal{P}$), il est possible de trianguler un ensemble de n points en $O(log(n))$.

4.2.3. Choix de $\mathcal{P}$

Le choix de l'ensemble de points $\mathcal{P}$ définissant les sommets de la triangulation s'effectue en deux étapes : une phase d'initialisation, générée au début du processus, et une phase d'affinement qui s'inscrit dans les itérations du processus de division, explicité dans le paragraphe suivant. La phase d'initialisation consiste quant à elle à placer aléatoirement n points sur l'image, selon une loi uniforme, puis à bruiter légèrement la position de ces points pour éviter la cocircularité de 4 points quelconques.

4.3. Phase de division

L'homogénéité de chaque triangle de la partition initiale est ensuite testée. Si le critère d'homogénéité d'un triangle T donné n'est pas satisfait, alors le barycentre du triangle est ajouté à $\mathcal{P}$, et une nouvelle triangulation est calculée par une itération de la méthode de Watson. Ce principe d'ajout de point est itéré jusqu'à convergence, *i.e.* jusqu'à vérification de l'homogénéité de tous les triangles de $DT(\mathcal{P})$.

L'homogénéité a été définie par la combinaison de deux critères. Le premier concerne les activités lues sur la région de l'image délimitée par T et consiste à calculer la moyenne (ou la médiane, moins sensible aux outliers) m et la variance σ^2 des niveaux de gris dans T, et à rechercher si une majorité des niveaux de gris des pixels de ce triangle (typiquement 95%) se trouve dans l'intervalle $[m-2\sigma;\ m-2\sigma]$. Le second critère est relatif à la taille des triangles et exprime le fait que $DT(\mathcal{P})$ ne présente pas d'hétérogénéité forte dans la forme des triangles. Un calcul de surface de T est donc effectué et T est rejeté si sa superficie est supérieure à une taille critique, évaluée par exemple comme la taille caractéristique d'un spot de l'image. Un triangle sera alors dit homogène s'il vérifie l'ensemble des critères.

D'autres critères peuvent également être envisagés par la suite, relatifs par exemple à des tests statistiques sur la distribution des niveaux de gris dans un triangle donné (monomodal pour un triangle homogène par exemple). De même, d'autres choix du point à ajouter dans T peuvent être étudiés, en fonction par exemple de la différence maximum d'activité dans le triangle.

4.4. Phase de fusion

A la convergence de la phase de division, tous les triangles de $DT(\mathcal{P})$ sont homogènes au sens des critères précédemment définis. L'étape de fusion consiste alors à réorganiser les triangles compte tenu de leur environnement, c'est-à-dire de leurs triangles adjacents.

Soient $T_j \in DT(\mathcal{P})$ et $\mathcal{T}$ l'ensemble des triangles adjacents à T_j. Un triangle $T \in \mathcal{T}$ est agrégé à T_j si les médianes et variances des niveaux de gris des triangles T_j et T sont semblables, par exemple au sens d'un test statistique (test non paramétrique de Kruskal Wallis pour les médianes) ou de simples seuils. A l'issue de l'étude du voisinage de T_j, l'ensemble des triangles agrégés forme un polygone étoilé, dont tous les points centraux sont éliminés de $\mathcal{P}$. La triangulation finale du domaine est alors finalement réalisée par application de l'algorithme de Watson sur cet ensemble de points de taille réduite. La triangulation $DT(\mathcal{P})$ obtenue après la fusion contient des triangles de grande taille sur les régions homogènes de l'image et d'autres de taille plus faible sur les zones de contours ou de bruit (speckle).

5. Résultats et discussions

5.1. Validation de méthodes

Nous proposons dans un premier temps d'utiliser le simulateur décrit dans le chapitre 1 pour évaluer quelques logiciels d'analyse d'images de puces couramment utilisés. Dans une première approche prospective, nous nous intéressons à deux outils disponibles au laboratoire, GenePix et Jaguar.

5.1.1. Description des logiciels

- GenePix

Genepix est distribué par Axon Instruments, Inc. Son interface simple permet d'identifier facilement les trois phases d'adressage, de segmentation et de quantification précédemment décrites. Le positionnement de la grille peut s'effectuer soit manuellement, par l'intermédiaire de quelques paramètres à saisir, ou par l'intermédiaire d'un fichier *.gal (Genepix Array List). Le bruit de fond peut être calculé de différentes manières. Par défaut, il est mesuré localement par la médiane des pixels inclus dans un masque.

L'utilisateur peut cependant choisir de le calculer globalement et ce de trois façons différentes : la moyenne des médianes du bruit de fond local de tous les spots ; la moyenne des moyennes du bruit de fond local de tous les spots ; une constante prédéfinie par l'utilisateur pour chacun des fluorochromes. La segmentation des spots est réalisée par un algorithme de cercle à diamètre adaptatif. Enfin, le logiciel permet de visualiser les données simplement par l'intermédiaire d'histogrammes et de graphiques et propose un grand nombre de résultats : plus de 70 paramètres statistiques peuvent en effet être affichés.

- Jaguar

Jaguar est un logiciel de la société Affymetrix, leader mondial dans le domaine des puces à ADN. Le fonctionnement de Jaguar nécessite le positionnement d'une grille sur chaque bloc de la puce de telle sorte qu'un et un seul spot soit présent par case. Ceci se fait automatiquement par l'intermédiaire d'un fichier CSV indiquant la position des blocs et des spots. Pour le bruit de fond, l'algorithme identifie les pixels sur les quatre côtés d'une case de la grille, puis calcule l'intensité médiane de ces pixels. Jaguar détermine un bruit de fond pour chacune des cellules de la grille. Concernant l'algorithme de segmentation de l'image, le logiciel donne le choix entre le cercle à diamètre constant et le cercle à diamètre adaptatif. Pour chacun des deux algorithmes, l'utilisateur peut choisir parmi deux modes de calcul pour estimer la valeur seuil permettant de détecter les spots sur la puce : un seuil faible *sf* et un seuil fort *sF*, où :

$sf = P_{75}Fond + (1.5 \times IQRFond)$

$P_{75}Fond$: percentile 75% des pixels du fond.

IQR_{Fond} : percentile 75%- percentile 25% du fond

et $sF = ((I90 - Fond)/2) + Fond$

I_{90} : percentile 90% des pixels du signal.

Fond : signal moyen du fond

Jaguar commence par identifier le pixel au centre de la case et détermine si l'intensité du pixel est supérieure à la valeur seuil du signal. Si l'intensité du pixel central dépasse la valeur seuil, le logiciel identifie tous les pixels qui lui sont contigus et dont l'intensité dépasse le seuil. Ces pixels sont utilisés pour définir le spot. Jaguar calcule la taille de la zone dans les deux directions x et y. Grâce à cette information, un cercle représentant le spot est alors défini. Si l'intensité du pixel central ne dépasse pas la valeur seuil, le logiciel recherche systématiquement dans la case un pixel dont l'intensité est supérieure à la valeur seuil. Si un tel pixel est trouvé, Jaguar procède de la même façon que précédemment pour tracer un cercle désignant le spot. Cependant, le cercle ne peut pas déborder sur la case voisine même si le spot le fait. Si aucun pixel ne dépasse la valeur seuil, le spot est dit manquant. L'intensité d'un spot est l'intensité médiane des pixels à l'intérieur du cercle. Pour l'estimation de l'erreur, le logiciel calcule la différence entre l'intensité de chaque pixel et l'intensité médiane puis retourne la valeur médiane de ces différences.

Jaguar possède des outils de visualisation de données performants tout comme GenePix mais à l'inverse de ce dernier, les résultats donnés par Jaguar sont peu nombreux : cinq pour chacun des fluorochromes plus un ratio des intensités.

Nous présentons une première étude prospective de la performance des logiciels précédents pour les étapes de segmentation et de quantification du processus d'analyse d'images. Pour ce faire, nous générons des images à l'aide du simulateur, pour lesquelles nous disposons de tous les paramètres (géométriques et de signal), qui servent donc dans la suite de données de référence auxquelles comparer les résultats.

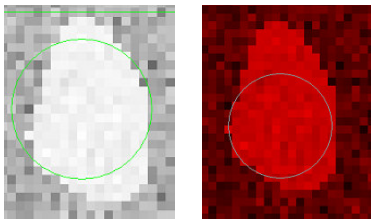
5.1.2. Etape de segmentation

Nous proposons dans un premier temps de tester les effets sur la segmentation des différentes déformations géométriques définies dans le chapitre 1, puis dans un second temps le comportement des logiciels en fonction du bruit dans l'image.

- Déformations géométriques

Nous avons tout d'abord testé le comportement des logiciels face à des spots déformés. Un unique spot de diverses images simulées a, pour ce faire, été déformé par l'intermédiaire du paramètre σ_d décrit dans le chapitre 1. L'algorithme de segmentation par cercle adaptatif a été utilisé dans les deux logiciels, pour pouvoir comparer les résultats.

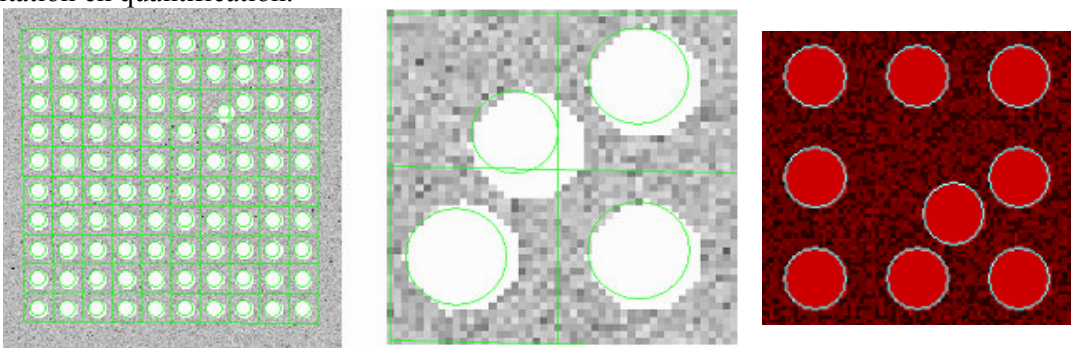
La Figure 2-12 présente un exemple de segmentation d'un spot déformé à l'aide des deux logiciels. Sur toutes les expériences effectuées, Jaguar agrège beaucoup plus de pixels du fond que GenePix, dont l'algorithme réussit à ne prendre en compte quasiment que des pixels faisant partie du signal. La statistique des mesures de Jaguar pour un spot s'en trouve alors complètement modifiée (diminution de la moyenne, augmentation de la variance) par rapport à la référence. Dans les deux cas cependant, une grande partie des pixels du spot a été oubliée (jusqu'à 45% de l'aire), ce qui fait prendre conscience des limites des méthodes de type cercle fixe ou adaptatif dans le cas de spots déformés. Dans ce cas précis, les méthodes de croissance de régions doivent par exemple donner de bien meilleurs résultats.



Jaguar GenePix

Figure 2-12 : segmentation d'un spot déformé

Nous nous sommes ensuite intéressés à la segmentation de spots déplacés. Pour ce faire, plusieurs images ont été générées, dans lesquelles un seul spot s'était déplacé de sa position théorique par l'intermédiaire du paramètre σ_t. Tous les autres spots de la puce sont par ailleurs supposés parfaitement circulaires, de même rayon et avec une activité constante à l'intérieur de chacun des spots, afin de juger les performances des algorithmes uniquement sur les spots déplacés. La Figure 2-13 présente un exemple de résultat d'une telle segmentation, sur une image de 1 bloc contenant 10 x 10 spots. Si GenePix parvient à segmenter correctement le spot (b), Jaguar ne segmente qu'une partie de ce dernier, contenue dans la case définie par la grille régulière calculée lors du plan de dépôt. Si le spot n'a qu'une faible intersection avec la case de la grille, l'échantillon des points du spot peut devenir statistiquement non significatif pour une exploitation en quantification.



a- Jaguar b- GenePix

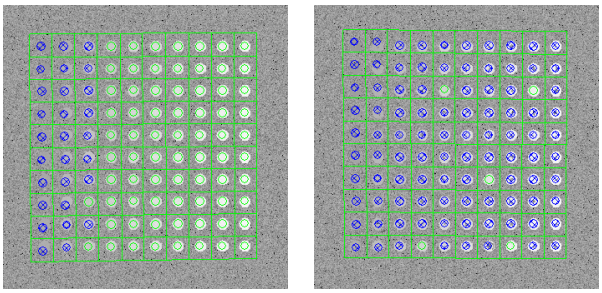
Figure 2-13 : segmentation d'un spot déplacé

Nous avons enfin examiné le cas de la segmentation de spots doughnut. Parmi toutes les simulations effectuées, nous présentons les résultats obtenus pour une image de 100 spots n'ayant aucune déformation avec une activité croissante en fonction du numéro du spot. Ainsi, le premier spot en haut à gauche possède l'activité minimale de la puce proche du niveau de bruit dans l'image alors que le spot en bas à

droite possède l'activité maximale. Nous avons crée une deuxième image ayant les mêmes propriétés géométriques que la première mis à part que tous les spots ont été définis comme doughnut. Ces deux images ont alors été segmentées à l'aide des deux logiciels : la Figure 2-14 présente le résultat obtenu à l'aide de Jaguar, et la Figure 2-15 avec Genepix, où :

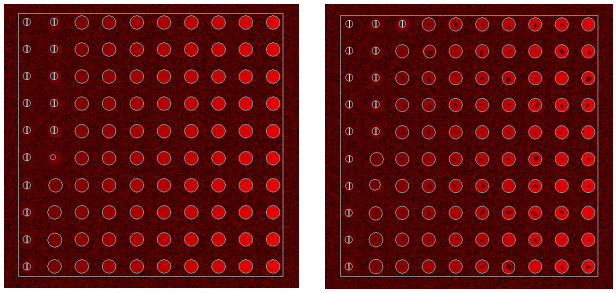
désigne les spots non segmentés correctement par Jaguar

et représente les spots non détectés par GenePix.



Cercles parfaits Doughnuts

Figure 2-14 : segmentation de spots doughnut – cas de Jaguar



Cercles parfaits Doughnuts

Figure 2-15 : segmentation de spots doughnut – cas de GenePix

L'algorithme de segmentation de Jaguar ne parvient pas à détecter les spots dès lors qu'ils possèdent ne serait-ce que quelques pixels de faible intensité et ce quelque soient la méthode (cercle fixe ou variable) et le seuil (faible ou fort) utilisés. Le fait que tous les spots sur la moitié gauche de la première image ne soient pas détectés vient du fait que leurs activités ne sont pas suffisamment contrastées par rapport au fond, point sur lequel nous reviendrons dans la partie consacrée à l'influence du bruit dans l'image.
La segmentation des deux mêmes images par l'algorithme de GenePix donne de biens meilleurs résultats : quelle que soit l'image, le nombre de spots détectés est quasiment équivalent : 85 % pour la première image et 84 % pour la seconde. Le fait que les spots situés sur la partie gauche des deux images ne soient pas détectés vient comme précédemment de leurs faibles intensités par rapport au bruit de fond.
Les valeurs des activités calculées par les logiciels pour des spots doughnut doivent être considérées avec attention, car les méthodes utilisées par les deux logiciels testés ne se concentrent que sur des formes circulaires de spots. Le « trou » d'activité caractéristique du doughnut fait alors changer la statistique de ces activités, ce qui peut être encore une fois préjudiciable à la bonne quantification des niveaux d'expression des gènes correspondants, et ce qui met une fois de plus en exergue la nécessité de développer en standard des méthodes plus génériques.

- Influence du bruit dans l'image

Puisque d'après Yang *et al*. [YANG00], le choix de la méthode d'estimation du signal du fond a autant d'importance que la précision de la segmentation des spots pour déterminer un niveau d'expression fiable des gènes, il nous a semblé important d'évaluer la performance des logiciels concernant la prise en compte du signal du fond. Pour cela, nous avons créé des séries d'images à spots circulaires, dont les

activités étaient croissantes en fonction de leur position dans l'image. Dans ces images, seul le niveau de bruit de fond variait. Nous avons alors observé les pourcentages de spots détectés par chacun des logiciels en fonction de ce niveau de bruit. La Figure 2-16 présente un exemple de résultat obtenu à partir d'une série d'images dont la Figure 2-17 montre un extrait.

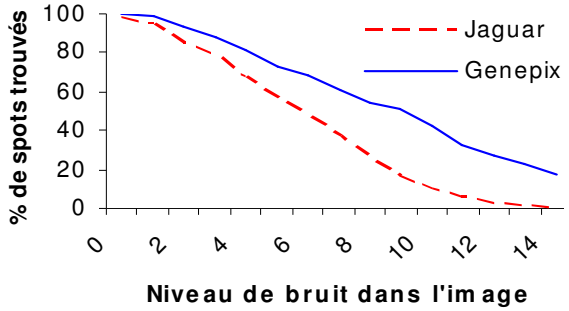


Figure 2-16 : évolution du nombre de spots segmentés en fonction du niveau de bruit

Ici, le niveau de bruit dans l'image est relié à la variance de la loi statistique chargée de modéliser le bruit de fond. L'algorithme de segmentation de GenePix détecte systématiquement sensiblement plus de spots que celui de Jaguar et cette différence ne fait qu'augmenter en fonction du niveau de bruit dans l'image. Ainsi, pour un niveau de bruit de 7 qui donne un bruit de fond important (Figure 2-17), Jaguar ne segmente que 37% des spots de l'image, GenePix en détectant quant à lui 64%.

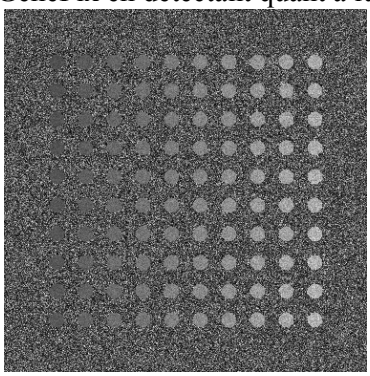
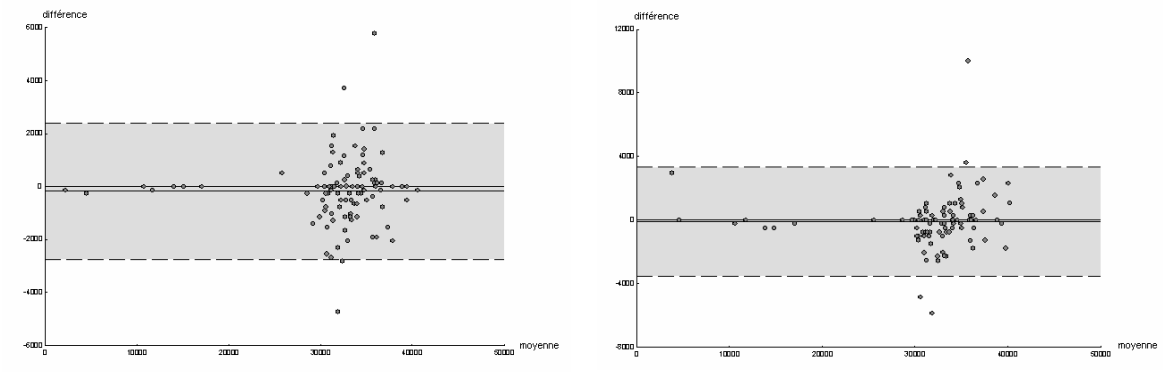


Figure 2-17 : exemple d'image simulée avec un bruit de fond important

5.1.3. Etape de quantification

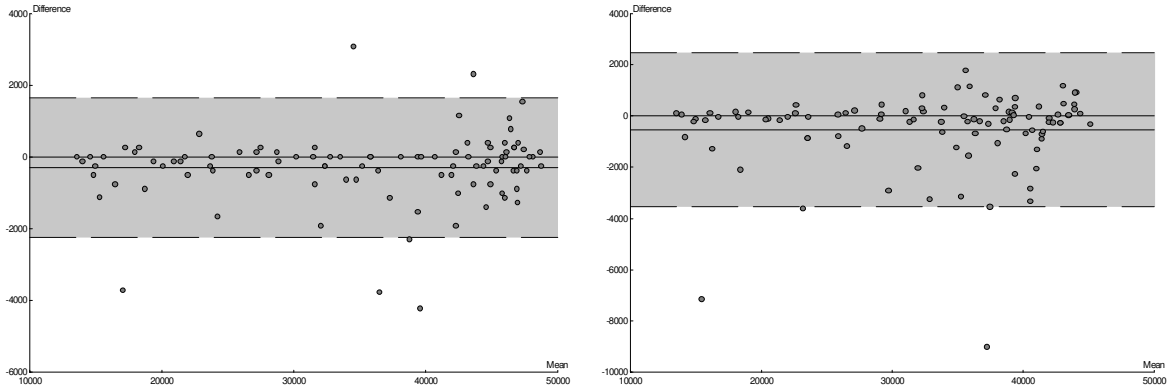
L'intensité de chaque pixel d'un spot est définie par une loi normale dont la moyenne est elle même calculée par une loi exponentielle et dont la variance vaut empiriquement un huitième de la moyenne [BISHOP74]. Nous nous intéressons ici au comportement de chacun des logiciels face à des spots non uniformes. Pour tester les logiciels dans ces conditions, nous avons créé plusieurs séries d'images en faisant uniquement varier la valeur du bruit de fond jusqu'à la valeur pour laquelle Jaguar ne détecte plus que 30% des spots de la puce. Dans une première série, les images ne contiennent que des spots parfaits, c'est à dire circulaires et sans trous. Dans une deuxième série, les images sont composées de spots parfaits et déformés. Enfin, une troisième série ne contient que des doughnuts. Pour chacune des séries, nous avons comparé les valeurs des médianes des activités calculées par chacun des logiciels avec les vraies valeurs calculées à la création de l'image, en utilisant un test de Bland et Altman [BLAND86].



GenePix Jaguar

Figure 2-18 : évaluation de la quantification des expressions par un test de Bland et Altman

S'il semble que les deux logiciels quantifient correctement les niveaux d'expression dans les spots parfaits (première série, Figure 2-18), GenePix est bien plus proche de la vraie quantification telle que proposée par le simulateur, l'intervalle de confiance étant 50% plus réduit que celui de Jaguar.
Dans le cas des déformations géométriques ou de doughnuts (deuxième et troisième séries), nous n'avons pas pu tester le logiciel Jaguar, celui-ci ne détectant en effet qu'un nombre très limité de spots sur chacune des images. Les résultats étant satisfaisants avec GenePix, nous avons poursuivi les tests et créé des images ayant une proportion équivalente de chaque déformation géométrique *i.e.* 1/3 de spots circulaires, 1/3 de spots déformés, 1/3 de dougnhuts, 30% des spots de l'image étant de plus déplacés. Nous avons également introduit une non-uniformité des spots croissante en fonction de leur intensité de départ. Nous avons alors comparé la médiane et la moyenne obtenues par GenePix et les valeurs simulées à la création de l'image (Figure 2-19).



Médianes Moyennes

Figure 2-19 : prise en compte des déformations dans la quantification

Dans les deux cas, une grande majorité des valeurs se situent dans la zone de confiance. Les points en dehors de cette zone correspondent aux doughnuts dont le diamètre intérieur est important. En ce qui concerne la répartition des points, nous ne pouvons pas tirer de conclusion sur le fait que plus les spots sont uniformes, plus les mesures sont correctes ou inversement. En effet, il existe une erreur systématique entre les valeurs mesurées par GenePix et les vraies valeurs. La zone de confiance du graphique concernant la médiane est enfin deux fois plus étroite que celle du graphique comparant la moyenne.

Bien que cette étude ne soit encore qu'à son début, quelques conclusions peuvent d'ores et déjà être tirées. Globalement, GenePix semble être le plus efficace des deux logiciels testés, que ce soit sur l'étape de segmentation ou dans la phase de quantification. Cependant, la trop grande simplicité des algorithmes utilisés, notamment concernant les techniques de segmentation, rend ces logiciels peu robustes dans des

cas qui se rencontrent finalement souvent sur de réelles images de puces à ADN (*e.g.* spots déformés). Ce qui nous conforte dans l'idée de proposer en standard des outils plus haut niveau permettant de réaliser par exemple une segmentation des spots efficace et la plus générique possible.

5.2. Segmentation par triangulation de Delaunay

Vingt expériences ont été réalisées à l'aide du simulateur, en faisant varier les divers paramètres géométriques et de signal. Pour chacune des images générées, l'algorithme de division/fusion a été appliqué et l'image résultante comparée à la carte théorique des contours des spots. Un exemple de résultat est présenté sur la Figure 2-20, pour lequel une image de paramètres géométriques (R_s=5µm, σ_r=0.05, S_H=10, S_V=10, H_B=15µm, H_V=15µm, B_H=1, B_V=1, P_{de}=0.1, P_{do}=0.2, P_{di}=0.1) et de signal (n=40, P_{du}=10^{-5}, A=10, m_r =3000, σ_r =20, α=0.15, P_{ue}=0.15, P_{oe}=0.15, T =parabolique, M=1) a été simulée.

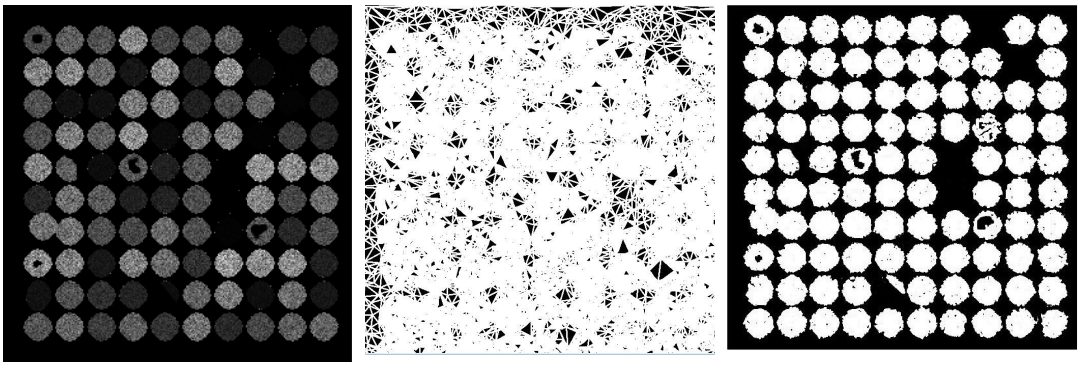


Image originale Triangulation finale Homogénéisation

Figure 2-20 : segmentation de spots par division/fusion

Pour évaluer la robustesse de la méthode, nous avons calculé pour chaque couple (référence, segmentation) le recouvrement spatial de ces deux images, en rapportant le nombre de pixels simultanément segmentés comme un même spot dans les deux images au nombre de pixels de spots dans l'image de référence (indice de confusion I). Les résultats moyens pour les 20 images simulées sont plus qu'encourageants, puisque environ 87% des pixels de spots dans toutes les images ont été identifiés comme tels (I = 0.87). Les erreurs d'étiquetage se situent dans leur grande majorité sur le contour des spots, puisque la phase d'homogénéisation construit ce dernier comme une ligne polygonale qui ne suit pas nécessairement le contour théorique des spots. De plus, pour quelques expériences où le bruit de poussière était important, quelques pixels ont faussement été classés comme spots.

A titre d'illustration, quelques unes des premières itérations de l'algorithme sont présentées sur la Figure 2-21.

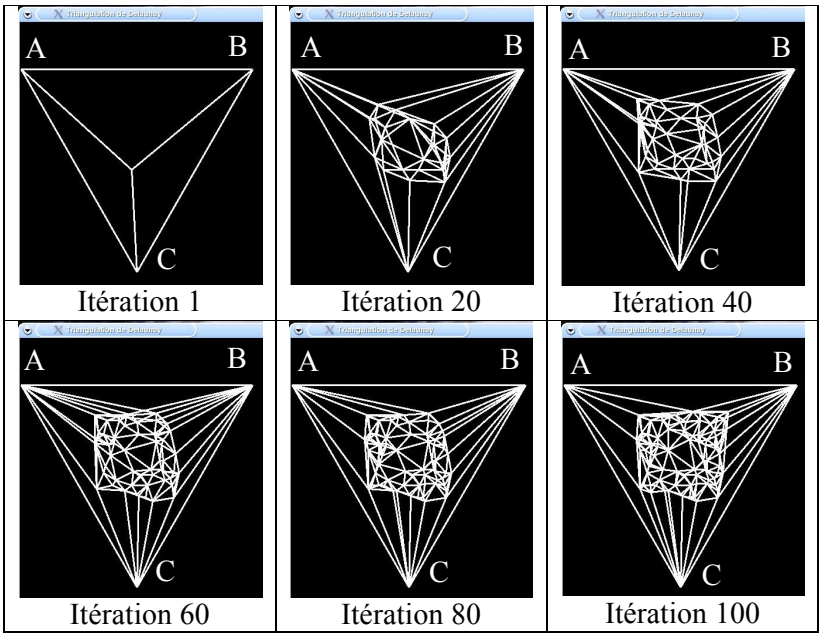
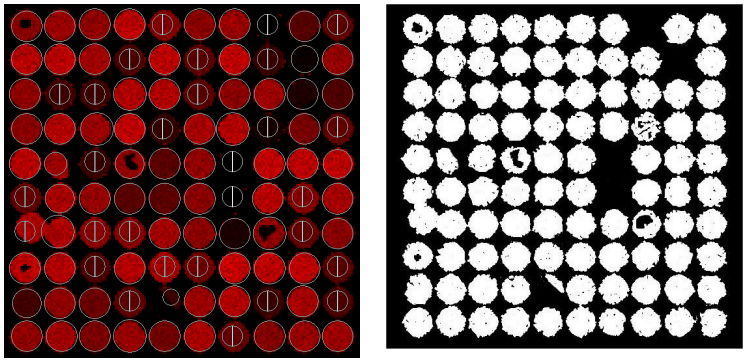


Figure 2-21 : évolution locale de la segmentation

Nous avons également comparé sur les mêmes images simulées les résultats de la segmentation fournie par les logiciels dont nous disposions au laboratoire (Jaguar et Genepix) avec ceux de la méthode de division/fusion. Le Figure 2-22 présente à titre d'illustration les résultats obtenus par Genepix sur l'image de la Figure 2-20. Vingt et un spots ne sont pas reconnus par ce logiciel (⌀) qui sont segmentés par l'algorithme proposé dans ce chapitre. Les spots manquants sont pour la plupart faiblement contrastés.



Segmentation GenePix Segmentation proposée

Figure 2-22 : comparaison de segmentations

Un des intérêts de la méthode de segmentation proposée ici est l'agrégation en une seule phase des étapes classiques d'adressage, de segmentation et de quantification. En effet, les structures de données sous-jacentes à l'algorithme donnent directement accès pour chaque spot à la liste des triangles le définissant, chacun d'entre eux portant une information fonctionnelle directement accessible. Il est donc également possible de réaliser des zones de confiance de quantification autour des spots, chaque zone étant maillée par un ensemble de triangles aux caractéristiques communes (par exemple au sens du critère d'homogénéité défini dans la méthode). L'adressage avant segmentation devient quant à lui obsolète, mais peut tout de même être réalisé, en calculant par exemple le dual de la triangulation de Delaunay (le diagramme de Voronoï) sur les centres de gravité des spots segmentés.
L'application de la méthode s'oriente maintenant vers la segmentation d'images réelles de puces à ADN.

Conclusion

La mise en évidence des carences et limites des méthodes classiques de segmentation des spots souligne l'importance de développer de nouveaux algorithmes plus robustes d'analyse d'images de puces, et de proposer ces derniers en standard sur des plateformes logicielles dédiées. Ce chapitre avait pour but d'une part de proposer une évaluation prospective des principaux outils proposés en standard, et d'autre part d'introduire une méthode de segmentation fondée sur une triangulation de Delaunay contrainte par le contenu de l'image, qui donne des résultats encourageants sur des simulations d'images de puces.

La continuité de ce travail inclut maintenant à court terme une reprise des tests des étapes d'analyse des images, sur un panel plus complet de logiciels et donc un ensemble plus étendu de méthodes. Ces tests permettront d'une part de proposer des domaines de validité pour les différents algorithmes mis en œuvre, et d'autre part des améliorations ponctuelles de ces derniers. A moyen terme, il s'agira alors de développer de nouvelles méthodes d'analyse (essentiellement de segmentation), à la lumière des résultats obtenus lors de l'étape de tests. L'ensemble devant concourir à une extraction de connaissances biologiques plus pertinente.

Chapitre 3 - Extraction de connaissances

Nous pouvons ajouter à nos connaissances, nous ne pouvons rien en retrancher.
Arthur Koestler (Les somnambules).

Introduction
L'extraction de connaissances à partir d'images de biopuces permet d'apporter des éléments de comparaison de l'état du transcriptome entre plusieurs conditions expérimentales données et/ou de suivre l'évolution du transcriptome en réaction à un stimulus expérimental. Trois problématiques majeures peuvent en particulier été identifiées, qui intéressent la médecine (association d'un phénotype à des gènes donnés à des fins diagnostiques), la biologie (compréhension de mécanismes génétiques responsables de l'apparition d'un phénotype) et le domaine sanitaire (détection de matériel génétique, par exemple pour l'identification d'agents pathogènes). Le contenu informationnel des données issues des biopuces est cependant directement fonction des méthodes mathématiques utilisées pour les étudier.
Ce chapitre présente alors les principales techniques utilisées dans ce domaine, et décrit leurs avantages et inconvénients. Deux nouvelles méthodes d'analyse sont ensuite introduites, ayant chacune un objectif particulier. La première, fondée sur une représentation des profils de gènes en courbes continues, et sur une classification de ces courbes, trouve son champ d'application dans l'étude de données temporelles. La seconde, fondée sur une représentation floue des données quantitatives, et sur un algorithme de fusion d'information modélisant des processus biologiques simples, se propose tout particulièrement d'étudier le gènes agissant au sein de mêmes réseaux de régulation. Ces deux méthodes sont validées sur des données de synthèse, puis appliquées sur des données réelles, et démontrent leur capacité d'extraire des informations pertinentes, confirmées par la littérature.

1. Normalisation

1.1. Pourquoi normaliser les données ?

Le chapitre 1 a permis de lister différents facteurs de variation affectant directement le niveau d'expression des gènes étudiés. Il est donc nécessaire de s'affranchir des facteurs de biais introduits lors de l'étape de préparation et d'acquisition pour une exploitation optimale des données. Trois objectifs sont plus précisément recherchés :

- s'assurer de la qualité des données pour une puce en particulier ;
- pouvoir comparer plusieurs puces utilisant le même ensemble de gènes et partageant une même condition expérimentale en double marquage ;
- pouvoir exploiter une puce quelconque possédant un gène ou un groupe de gènes d'intérêt (récupération de données publiques par exemple).

Dans ce cadre, la normalisation est définie comme un processus de transformation des intensités, précédant toute analyse des données, et permettant l'étude effective des niveaux d'expression des gènes.

1.2. Que normaliser ?

Le plus souvent, la correction apportée dans le cas d'un double marquage est effectuée sur le ratio des intensités brutes des signaux issus des deux canaux d'acquisition. Ce choix offre pour avantages de manipuler directement le concept de modulation d'expression qui lie les deux canaux et de s'affranchir de certains facteurs de biais multiplicatifs constants.

1.3. Comment normaliser ?

La normalisation des données s'effectue en deux étapes : une première d'estimation de la qualité des spots, et une seconde de normalisation à proprement parler, qui consiste à corriger les valeurs des données. De nombreuses approches ont aujourd'hui été développées, qui restent cependant encore à évaluer :

- normalisation par décalage de la moyenne ;
- recalage de la moyenne des ratios à 0 par addition d'un facteur, constant ou non, basé sur l'hypothèse d'une distribution normale parfaite ;
- normalisation par intensité globale ;
- normalisation par la variance: en plus de la normalisation par la moyenne, la variance est ajustée à 1 pour tous les spots ;
- réduction de l'importance de la dispersion des données pour les faibles intensités en prenant en compte les Z-scores.

Yang *et al.* [YANG01] et Schuchhardt *et al.* [SCHUCHHAR00] ont comparé de nombreuses méthodes de normalisation et concluent qu'aucune n'est universelle, chacune possédant ses avantages et inconvénients en fonction du cas particulier expérimental étudié.
Plutôt que de baser la normalisation sur les propriétés (hypothétiques) de la distribution des données, il est encore possible de les comparer à un même étalon interne :

- soit des gènes connus pour leur absence de modulation (ou présumés comme tels) dans les conditions expérimentales testées ;
- soit des séquences artificielles (exogènes) apportées en quantités connues avec les ARN messagers testés ;
- soit des contrôles universels (normalisation idéale, mais réalisation extrêmement difficile).

1.4. Que faire après la normalisation ?

Une fois les données pertinentes normalisées, la dernière question à se poser avant l'analyse effective de l'expérience concerne la valeur des ratios minimale au-delà de laquelle la modulation peut être considérée comme pertinente. Le signal d'un spot est par exemple considéré comme pertinent si $(t-b)>c.d$, où t est l'intensité du spot, b le bruit de fond local, d l'écart type de la différence entre bruit de fond local et bruit de fond intrinsèque, et c un seuil d'élimination. Ce dernier paramètre conditionne bien évidemment le résultat du filtrage (un seuil élevé permettra de ne garder que les spots les plus fiables, mais en nombre peu élevé). Bien que cette étape soit encore essentiellement effectuée de manière empirique (*e.g.* c=5 dans [DERISI00], c=2 dans [MARC02]), il est maintenant envisagé d'utiliser des outils statistiques pour obtenir une valeur de seuil de ratio plus robuste, ce qui nécessite alors de réaliser plusieurs fois l'expérience pour obtenir des effectifs suffisants [LEE00].

2. Méthodes d'analyse de données d'expression de gènes

Les données de départ sont donc maintenant sous la forme d'une matrice $X=(x_{ij})$ d'expression de gènes, $1 \le i \le N$ et $1 \le j \le p$, n étant le nombre de gènes pertinents étudiés selon les p expériences. x_{ij} est donc le niveau d'expression du gène i sous la condition j. La question posée ici concerne le regroupement de gènes ayant des ensembles de niveaux d'expression, *i.e.* des profils, semblables, en K classes distinctes. L'idée sous-jacente est que si les gènes possèdent la même régulation transcriptionnelle, ils ont des chances d'être impliqués dans la même réponse et éventuellement d'être régulés par le même mécanisme. Cela permet de décomposer le processus en séries d'événements impliquant des groupes de gènes et de voir plus facilement les composantes de la réponse au phénomène étudié (pathologie, stimuli…) Cela peut également permettre d'attribuer une fonction putative à des gènes n'ayant jamais été étudiés. Golub *et al.* [GOLUB99] nomment ces deux problématiques par class prediction et class discovery, ce qui d'un point de vue reconnaissance des formes correspond aux problèmes bien connus de discrimination et de classification.

2.1. Discrimination

Dans le cadre de la discrimination, un ensemble d'apprentissage permet de déterminer K classes *a priori*, et l'objectif est d'expliquer et de prédire pour un nouveau gène son appartenance à l'une des K classes.

2.1.1. Analyse discriminante linéaire et quadratique

L'analyse discriminante linéaire (resp. quadratique) part de la connaissance de la partition en K classes des gènes x_i de la puce et cherche les combinaisons linéaires (resp. quadratiques) des variables x_{ij} qui conduisent à la meilleure discrimination entre les classes (par exemple qui maximisent la variabilité inter classes). Appliquée à l'étude des cellules B, de la leucémie et à des lignées de cellules tumorales [DUDOIT02], l'analyse discriminante linéaire s'est en particulier révélée très efficace pour la différenciation de ces types de cellules.

2.1.2. Modèles statistiques

Dans ce type de méthodes, certaines hypothèses sur la représentation des lois de distributions des classes de gènes sont effectuées, souvent sous la forme de lois paramétriques, dont les paramètres sont déterminés à l'aide de l'ensemble d'apprentissage. Une règle de décision type règle de Bayes [BARASH02] permet ensuite de classer tout gène candidat dans l'une des K classes prédéfinies.

2.1.3. Méthodes des plus proches voisins

Les méthodes des plus proches voisins sont fondées sur une distance entre paires d'observations. La règle classique des k plus proches voisins se propose de classer un nouveau gène x tout d'abord en recherchant dans l'ensemble d'apprentissage les k gènes les plus proches de x, puis en prédisant la classe de x en fonction de la classe la plus représentée parmi ses k voisins. Bien évidemment, ce type de méthode dépend fortement du choix de k, de la métrique utilisée et de l'ensemble d'apprentissage.

2.1.4. Support Vector Machines

Introduit par Vapnik dans les années 70 [VAPNIK98], les techniques de Support Vector Machines (SVM) sont des méthodes supervisées de classification, très robustes vis à vis de données éparses et bruitées. Le principe général est une séparation d'un ensemble de données labellisées (ensemble de test) par un hyperplan maximisant la distance aux points de test. Dans le cas où aucune séparation par un tel hyperplan n'est possible, il y a possibilité de coopération entre SVM et une technique de noyaux qui réalise une séparation non linéaire. Les données dont on ne connaît pas l'étiquetage sont ensuite testées par rapport à l'hyperplan séparateur, et donc classées (avec un intervalle de confiance dépendant par exemple de la distance à l'hyperplan).
Ce type de méthodes a récemment beaucoup été utilisé dans le domaine des puces à ADN, que ce soit pour la classification de gènes ([BROWN97] [PAVLIDIS01]), la classification de tissus ([MUKHERJEE99] [FUREY00]) ou l'étude de pathologies ([GUYON01]).

Dudoit *et al*. [DUDOIT02] ont proposé une étude très complète sur les méthodes de discrimination, et concluent que des classifieurs simples tels que la méthode du plus proche voisin donnent des résultats satisfaisants en comparaison de techniques plus complexes.

2.2. Classification

Aucun ensemble d'apprentissage n'est ici nécessaire, les K classes étant automatiquement générées et la seule intervention extérieure résidant dans l'identification de la partition calculée. Les méthodes de classification semblent aujourd'hui plus utilisées que les techniques de discrimination, en particulier en raison de l'historique des méthodes d'analyse d'expression des gènes, débuté en 1998 par M. Eisen et son équipe de l'université de Californie [EISEN98].

2.2.1. Classification hiérarchique

Ce type de classification repose sur la construction hiérarchique d'un arbre (dendogramme). Partant d'une matrice de distances entre gènes, la première étape consiste à sélectionner les deux gènes les plus proches au sens d'une métrique choisie. Un gène moyen est alors calculé, et réinjecté dans la population à la place de ses deux gènes générateurs. Le processus est itéré jusqu'à l'agrégation du dernier gène isolé. L'algorithme dépend donc d'une distance (euclidienne, de Manhattan, de Mahalanobis,…) et d'une stratégie d'agrégation des branches (lien simple, lien complet, UPGMA, WPGMA, méthode de Ward) dont le choix est crucial pour une bonne classification des gènes. Une fois le dendogramme calculé, ce dernier peut être décomposé en un nombre de classes K simplement en coupant les $K-1$ branches les plus

significatives de l'arbre. Le choix de K est essentiellement empirique ou repose sur des critères statistiques (critères d'Akaike).
Eisen *et al.* [EISEN98] proposent la première utilisation de cette méthode pour l'analyse du transcriptome, méthode de nombreuses fois reprise depuis, par exemple dans [ALIZADEH00] pour l'analyse de données de lymphomes chez l'humain, ou dans [CHU98] pour l'étude des données de sporulation chez la levure du boulanger.

2.2.2. Partitionnement par k-moyennes

La classification par partitionnement est différente de la classification hiérarchique, puisqu'elle suppose connu *a priori* le nombre de classes K. La méthode des k-moyennes est une approche itérative de calcul de K centres de classe c_l et d'affectation des individus aux classes selon un critère de distance. Si cette méthode est simple de mise en œuvre et d'utilisation, elle n'assure cependant pas la convergence vers la meilleure partition possible. Si l'algorithme de base a peu été appliqué en classification de données issues de puces à ADN [TAVAZOIE99], des variantes déterminant incrémentalement le nombre K de classes optimales ont été proposées [BENDOR99] [HERWIG99].

2.2.3. Cartes de Kohonen

Les cartes de Kohonen (ou Self Organizing Maps - SOM) [KOHONEN97], issues de la littérature sur les réseaux de neurones artificiels, représentent une alternative séduisante aux méthodes précédentes. Le principe de base est similaire à celui des k-moyennes, l'originalité résidant dans la façon dont les positions des centres de classe sont réactualisées. L'idée est d'introduire un espace de représentation à une ou deux dimensions, représentant les notions de voisinage entre classes. Si un gène x_i est affecté à une classe k, la position des centres de classe c_l est réactualisée pour tenir compte de la nouvelle affectation de x_i par :

$$c_l(t+1) = c_l(t) + h_{kl}(t).(x_i - c_l(t)),\ 1 \leq l \leq K$$

La fonction $h_{kl}(t)$ doit être non nulle si et seulement si les classes k et l sont suffisamment proches dans l'espace de représentation, et tendre vers 0 lorsque t tend vers l'infini.
Des données de *Saccharomyces cerevisiae* [HERRERO01] [TAMAYO99] [TÖRÖNEN99] et humaines [TAMAYO99] ont par exemple été étudiées en utilisant ce type de méthode.

2.2.4. Méthodes de bi-clustering

Toutes ces méthodes de classification permettent d'approcher le rôle des gènes dans une réponse, mais peu prennent en compte le fait que les gènes agissent au sein de réseaux de régulation. Un état phénotypique, qui est la résultante de l'action de nombreux gènes aux profils d'expression éventuellement dissemblables, ne sera pas mis en valeur puisque les méthodes précédentes ne pourront pas associer les gènes correspondants. Il faut alors envisager de rechercher les gènes dont les profils d'expression sont spécifiques d'un phénotype d'intérêt, et donc envisager deux phases d'analyse : une première étape de classification sur les expériences et sur les gènes, puis une exploitation automatisée des connaissances biologiques par des méthodes de data mining. Le but étant alors la définition de groupes de gènes biologiquement cohérents et signant une fonction physiologique particulière. Les méthodes de bi-clustering proposent de réaliser les première étape, et quelques résultats ont d'ores et déjà été obtenus sur des données du cancer du colon et de la leucémie [GETZ00] ou sur des données de la levure du boulanger et des cellules B humaines [CHENG00].

2.2.5. Modèles de mélange

Ce type d'approche s'appuie sur l'hypothèse que chaque classe de gènes peut être représentée par une loi statistique donnée (souvent normale multivariée). Cette distribution est caractérisée par un ensemble de paramètres, estimés par une méthode de type maximum de vraisemblance. La classification est alors réalisée par exemple suivant le principe du maximum *a posteriori*. Ce type de méthodes est très souple (de nombreux modèles de mélange sont possibles), repose sur un cadre théorique probabiliste bien établi,

et c'est pourquoi de nombreux auteurs l'ont utilisé dans ce domaine, notamment dans le cadre de données répétées [GHOSH02] [MCLACHLAN02] [YEUNG01].

2.2.6. Réduction de dimension

Le but des méthodes de réduction de dimension est de résumer et de hiérarchiser l'information contenue dans la matrice *X* des données d'expression. La méthode la plus utilisée dans ce cadre est l'Analyse en Composantes Principales (ACP), qui induit un changement de représentation permettant de modéliser de façon plus synthétique, et sans les dénaturer, les variations des données. La méthode consiste tout d'abord à standardiser les données, puis à trouver les éléments propres de la matrice de corrélation entre les gènes. Les gènes sont alors projetés sur les axes propres principaux donnant accès le plus souvent à une représentation graphique 2D ou 3D des variations des données initiales. Raychaudhuri *et al*. [RAYCHAUDH00] ont par exemple utilisé l'ACP sur les données de sporulation de la levure *Saccharomyces cerevisiae* et comparé leurs résultats aux données de Chu *et al*. [CHU98], dans lesquelles étaient identifiées sept classes. De même, Crescenzi *et al*. [CRESCENZI01] ont identifié par ACP cinq composantes indépendantes permettant de souligner des informations pertinentes pour la classification de lignées tumorales.

Un algorithme de calcul de l'ACP, la décomposition en valeurs singulières, est également utilisé sur des matrices d'expression de gènes [ALTER03], de même que quelques approches de même philosophie que l'ACP. Parmi celles-ci, citons ici le Gene Shaving [HASTIE00], s'intéressant aux sous-ensembles de gènes fortement corrélés et dont la moyenne a une forte variance, ou l'Analyse en Composantes Indépendantes [LIAO02], complément de l'ACP cherchant les expressions des gènes comme combinaisons linéaires de sources statistiquement indépendantes.

Notons enfin que certains auteurs [YEUNG00-B] affirment de l'ACP est peu désignée dans le cas de données manquantes.

2.3. Autres approches

2.3.1. Logique floue

Il s'agit dans ces méthodes d'interpréter les niveaux d'expression à l'aide d'une approche floue, qui permet de trouver par exemple des relations entre gènes. Globalement, chaque niveau d'expression ou groupe de niveaux lu est relié à un quantificateur flou (« grand », « petit », « moyen »), et un ensemble de règles floues est appliqué pour inférer un résultat (par exemple « si l'expression de G1 est grande et l'expression de G2 est faible, alors G3 est faiblement inhibé). Quelques auteurs utilisent cette approche, intégrée à des réseaux de neurones flous [AZUAJE01] [TOMIDA01], des algorithmes de classification [GUTHKE01] ou encore des règles heuristiques de décision [WOOLF00], le principal argument étant la possibilité de prendre en compte les ambiguïtés des mesures (imprécisions dues au bruit lors de la mesure, imprécision lors de la détection des spots sur la plaque…) dans le processus de quantification.

2.3.2. Théorie des graphes

Il est possible d'associer à la matrice *X* une matrice de similarité et une structure de graphe permettant alors de traiter le problème de classification de gènes à l'aide des outils issus de la théorie des graphes (*e.g.* coupe minimale, recherche de cliques maximales dans le graphe). Le regroupement des gènes devient alors par exemple une recherche de sous graphes fortement connectés [BENDOR99] [HARTUV99] [XING01]

2.3.3. Approches bases de données

Ce type d'approche (data mining) permet de trouver des relations de dépendances entre les niveaux d'expression des différents gènes de *X* à l'aide d'outils de base de données. Il peut s'agir par exemple de définir un réseau de régulation de gènes par un modèle simple d'activateurs/inhibiteurs et des règles d'association [KOTALA01] [LEMKIN00] [TUZHILIN02], ou de rechercher des dépendances fonctionnelles dans la matrice X [AUSSEM02].

Nous apportons ici une contribution sous la forme de deux méthodes, l'une s'intéressant à une problématique de classification de profils de gènes semblables, l'autre se focalisant plus sur la recherche de groupes de gènes participant par exemple à un même réseau de régulation.

3. Méthodes proposées

3.1. Analyse en composantes principales fonctionnelles

L'analyse en composantes principales fonctionnelles (ACPF) est issue des travaux de Ramsay et Silverman [RAMSAY97] sur la description des modes de variations d'un ensemble de courbes, et trouve ses applications dans des domaines variés allant de la modélisation de la forme d'os à la croissance humaine, en passant par le traitement d'images échocardiographiques ou la reconnaissance d'écritures manuscrites. Puisque l'ACP est classiquement utilisée en analyse d'expression de gènes, il nous a paru naturel d'étendre cette technique en considérant que les profils d'expression de gènes définissent une courbe continue $x_i(s)$, la matrice X devenant alors un ensemble de courbes dont il faut analyser les principaux modes de variation [BARRA04-A].

3.1.1. Principe

Les données fonctionnelles sont un ensemble de courbes continues $(x_i(s), i=1..N), s \in [a,b]$. X est ainsi composé de N individus et de la variable continue s (dimension infinie). Le problème consiste alors à trouver un ensemble de fonctions $\varepsilon_\alpha(s)$, α dans $1..q$, mutuellement orthogonales, de norme 1, chacune d'entre elles maximisant :

$$\frac{1}{N}\sum_{i=1}^{N}\left(\int_a^b \varepsilon_\alpha(s)x_i(s)ds\right)^2 = \frac{1}{N}\sum_{i=1}^{N} f_i^2$$

On introduit une fonction de covariance $v(s,t) = \frac{1}{N}\sum_{i=1}^{N} x_i(s)x_i(t) = \frac{1}{N}\mathbf{X}(s)\mathbf{X}^T(t)$

Et le problème de maximisation peut se mettre sous la forme :

$$\frac{1}{N}\sum_{i=1}^{N}\int_a^b \varepsilon_\alpha(s)x_i(s)ds\int_a^b \varepsilon_\alpha(t)x_i(t)dt = \int_a^b\int_a^b \varepsilon_\alpha(s)\varepsilon_\alpha(t)\left(\frac{1}{N}\sum_{i=1}^{N} x_i(s)x_i(t)\right)dsdt$$

$$= \int_a^b \varepsilon_\alpha(s)\left[\int_a^b v(s,t)\varepsilon_\alpha(t)dt\right]ds$$

Posons

$$V\varepsilon_\alpha(s) = \int_a^b v(s,t)\varepsilon_\alpha(t)dt = \langle v(s.,), \varepsilon_\alpha\rangle \qquad (1)$$

La maximisation se réduit donc à $Max \langle \varepsilon_\alpha, V\varepsilon_\alpha\rangle$, sous $||\varepsilon_\alpha|| = 1$. Posons alors

$$\frac{1}{N}\sum_{i=1}^{N} f_i^2 = \lambda = \lambda\int_a^b \varepsilon_\alpha(s)\varepsilon_\alpha(s)ds$$

On a :

$$\int_a^b \varepsilon_\alpha(s)V\varepsilon_\alpha(s)ds = \lambda\int_a^b \varepsilon_\alpha(s)\varepsilon_\alpha(s)ds \text{ et donc } V\varepsilon_\alpha(s) = \lambda\,\varepsilon_\alpha(s).$$

La maximisation de

$$\frac{1}{N}\sum_{i=1}^{N} f_i^2 = \lambda$$

est alors équivalente à la recherche des éléments propres de la matrice V, appelée opérateur de covariance. C'est le principe de l'ACP fonctionnelle.

3.1.2. Méthode de calcul

La résolution du problème d'éléments propres consiste alors tout d'abord à centrer les données fonctionnelles, puis à développer ces dernières en des combinaisons linéaires de fonctions de base. Chaque composante principale est alors représentée par une fonction continue.

Supposons que chaque fonction continue $x_i(s)$ est combinaison linéaire de K fonctions de base $(\phi_k(s))_{k=1..K}$ pondérées par des coefficients c_{ik} regroupés dans une matrice C $(N*K)$:

$$X(s) = C\Theta(s).$$

La fonction de covariance $v(s,t)$ devient alors :

$$v(s,t) = \frac{1}{N} X^T(s)\, X(t) = \frac{1}{N} \Theta(s)^T\, C^T C \Theta(t)$$

Soit W la matrice $K*K$ de coefficients $w_{kl} = \int \Phi_k(s)\Phi_l(s)ds$

Si la fonction propre $\xi(s)$ s'écrit dans la base des $\phi_k(s)$: $\xi(s) = \Theta(s)^T.b$, où $b = [b_1 \cdots b_k]^T$ alors par définition de V (1) :

$$V\xi(s) = \int_a^b v(s,t)\xi(t)dt = \frac{1}{N} \int \Theta^{\mathrm{T}}(s)C^{\mathrm{T}}C\Theta(t)\Theta^{\mathrm{T}}(t)bdt = \frac{1}{N} \Theta^T(s)C^TCWb.$$

Et

$$\frac{1}{N} \Theta^T(s)C^TCWb = \lambda\xi(s) = \lambda\Theta(s)^T.b$$

Il existe donc une équation matricielle pour tous les s telle que :

$$\frac{1}{N}\ C^TCWb = \lambda b \qquad (2)$$

Compte tenu du fait que $\int \xi^2(s)ds = 1$ on a :

$\int b^T\Theta(s)\Theta(s)^T bds = b^TWb = 1$

Soit u un vecteur normalisé tel que $u = W^{1/2}b$. Alors (2) devient :

$$\frac{1}{N}\ (W^{1/2}C^TC\ W^{1/2})u = \lambda u.$$

Si la base est orthogonale, alors $W=I$ et $\frac{1}{N}\ C^TC\ u = \lambda u$, et la méthode consiste donc alors à :

- Calculer les m éléments propres $\{(u_1 \ldots u_m), (\lambda_1 \ldots \lambda_m)\}$ de $Q = \frac{1}{N}\ (W^{1/2}C^TC\ W^{1/2})$
- Calculer les m composantes principales $\xi_i(s)$, $i=1..m$ par $\xi(s) = \Theta(s)^T.b$.
- Calculer comme en ACP classique la contribution de chaque composante principale par le rapport entre la valeur propre correspondante et la somme des valeurs propres.
- Calculer la position de chaque individu (ou courbe $x_i(s)$) dans le système des composantes principales par le produit $X\varepsilon$, $\varepsilon = [\xi_1 \ldots \xi_m]$. Si $U = [u_1 \ldots u_m]$, $B = [b_1 \ldots b_m]$, alors par définition $F = X\varepsilon = (C\Theta)(\Theta TB) = CWB = CWW\text{-}1/2B = CW1/2U$.
- Les scores de la courbe $x_i(s)$ dans le système des m composantes principales sont définies sur la $i^{\text{ème}}$ ligne de F. Comme en ACP classique, on se limite à m=2 ou m=3. L'éloignement d'un point par rapport à l'origine traduit l'écart de la courbe considérée par rapport à la courbe moyenne. Le scalaire f_{ij} est le score de l'individu i par rapport à la composante principale j.

3.2. Une utilisation de la fusion de données

La connaissance des méthodes de fusion d'informations, développées par ailleurs en imagerie cérébrale, amène tout naturellement à appliquer ce type de principe dans le cadre de l'extraction de connaissances des images de puces à ADN, lorsqu'il s'agit par exemple de trouver des relations simples entre gènes inhibiteurs et activateurs. Dans la suite, nous proposons une étude de faisabilité concernant l'application sur l'analyse des données d'expression de gènes d'une partie des concepts développés dans la partie 1 [BARRA02-A].

Nous supposons sans restriction ici qu'à l'issue de la normalisation, les niveaux d'expression des gènes x_{ij} sont dans l'intervalle [-1,1]. Nous proposons alors de trouver les gènes de X activés par un sous-ensemble A de gènes et inhibés par un sous-ensemble I d'autres gènes, *i.e.* trouver les gènes C satisfaisant :

$$(\forall j \in [1,p])\ x_{C,j} = \sum_{a \in A} \alpha_a . x_{aj} - \sum_{i \in I} \alpha_i . x_{ij}, \quad \sum_{k \in \{a,i\}} \alpha_k = 1,\ \alpha_i,\ \alpha_k > 0 \qquad (3)$$

Notre méthodologie se fonde sur trois étapes :

- modélisation des données concernant les gènes des ensembles A et I,
- combinaison des informations et création d'un gène prototype, modélisant les interactions entre activateurs et inhibiteurs,
- décision : recherche dans X des gènes les plus proches de ce prototype.

Pour illustrer la méthode, nous supposerons dans la suite chercher l'ensemble des gènes activés par un gène A et inhibés par un autre gène I (i.e. $|A|=|I|=1$). Si A est la ligne de X où s'exprime A, donnée par les coefficients x_{Aj}, et I la ligne de X où s'exprime I, s'écrit alors :

$$(\forall j \in [1,p])\ x_{Cj} = \alpha x_{Aj} - \beta x_{Ij}\ \alpha \geq 0;\ \beta \geq 0,\ \alpha + \beta = 1 \qquad (4)$$

3.2.1. Modélisation des données

Les données sont deux vecteurs $A = (x_{A,j})$ et $I = (x_{I,j})$, $j \in [1,p]$, de données réelles dans [-1,1]. Ces valeurs reflètent l'intensité de l'expression de chacun des deux gènes sous les différentes expériences, qu'elle soit "forte" (pour des valeurs proches de 1), "moyenne" ou "faible". Nous attribuons donc à chaque valeur un qualificateur flou correspondant à ces notions. L'intervalle des valeurs possibles des x_{ij} est donc tout d'abord décomposé selon une partition floue ($E_1, E_2 \ldots E_q$) à l'aide d'ensembles flous, définis par leurs fonctions d'appartenance ($\mu_1, \mu_2 \ldots \mu_q$) telles que :

$$(\forall x \in [0,1]) \sum_{i=1}^{q} \mu_i(x) = 1 \qquad (5)$$

Par exemple, [-1,1] peut être classiquement partitionné (q=5) par cinq qualificateurs linguistiques "très faible"(tf), "faible"(f), "moyen"(m), "fort"(F) et "très fort"(TF). Le choix q=5 a été considéré dans la suite pour prendre en compte la largeur de l'intervalle [-1,1] et proposer une gamme de variations suffisantes de comportements (Figure 2-23).

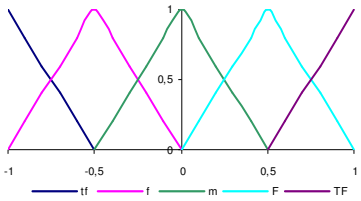


Figure 2-23 : exemple de quantificateurs flous

Chaque valeur $x_{.j}$ se verra donc attribuer un quintuplet de réels dans [0,1] (*tf, f, m, F, TF*) à partir de cette partition, avec $tf+f+m+F+TF=1$, correspondant aux degrés d'appartenance du réel $x_{.j}$ aux ensembles flous précédemment décrits. Dans un premier temps, $x_{.j}$ se verra attribué le qualificateur flou pour lequel

il a le plus grand degré d'appartenance. Dans un second temps, que nous souhaitons entreprendre ultérieurement, les quintuplets seront intégralement pris en compte dans un système de règles floues pour modéliser le comportement d'un système activateurs/inhibiteurs.

3.2.2. Combinaison des informations

Les valeurs $x_{A,j}$ et $x_{I,j}$ de chaque expérience j doivent maintenant être combinées pour déterminer une valeur de l'expression du gène C résultant de la même expérience. Nous nous inspirons des techniques de fusion de données pour calculer cette combinaison. Nous proposons ici un opérateur de combinaison (*i.e.* une fonction F : [-1,1]×[-1,1] → [-1,1], qui au couple $(x_{A,j},x_{I,j})$ associe la valeur de l'expression $x_{C,j}$ issue des opérateurs d'agrégation à base de règles floues. I.Bloch [BLOCH96] propose une classification des opérateurs de fusion, non seulement selon leur caractère sévère (conjonctif) ou indulgent (disjonctif) mais aussi par rapport à leur comportement vis à vis du contexte. Trois classes sont alors définies :

- les opérateurs indépendants du contexte et à comportement constant ;
- les opérateurs indépendants du contexte et à comportement variable ;
- les opérateurs dépendant du contexte.

Le choix qui semble le plus naturel ici est celui d'un opérateur de combinaison dépendant du contexte, puisque l'intensité relative des expressions des activateurs et des inhibiteurs va conditionner l'expression du gène C. Nous proposons un tel opérateur, fondé sur des règles simples correspondant au comportement intuitif du modèle activateur/inhibiteur. (*e.g.* si l'activité de A est faible et celle de I forte, alors celle de C est faible). Nous notons dans la suite Max(x,y) (resp. Min(x,y)) l'union (resp. l'intersection) logique. Ces deux opérateurs correspondent respectivement au plus grand des opérateurs conjonctif et au plus petit des opérateurs disjonctifs. Un opérateur F réalisant la combinaison des expressions des gènes A et I peut alors par exemple s'écrire :

x_{Aj} \ x_{Ij}	tf	f	M	F	TF
tf	Mean	Min	Min	Min	Min
f	Max	Min	Min	Min	Min
m	Max	Max	Max	Min	Min
F	Max	Max	Max	Min/3	Min/3
TF	Max	Max	Max	Min/3	Min/4

Tableau 2-4 : opérateur de combinaison

Ainsi, si par exemple $x_{A,j}$ est faible et $x_{I,j}$ est fort, l'expression du gène C est fortement inhibée par I et $x_{C,j}=Min(x_{A,j}, x_{I,j})$. De même, si l'activateur et l'inhibiteur s'expriment très fortement, alors le gène résultant aura une expression moyenne. Puisque A et I ont leur plus grand degré d'appartenance dans l'ensemble flou TF, cela signifie que $0.75 \leq x_{A,j} \leq 1$ et $0.75 \leq x_{I,j} \leq 1$. Nous ramenons donc $x_{C,j}$ dans l'ensemble m (moyen) en divisant le minimum de ces deux valeurs par 4, assurant ainsi que $0.18 \leq x_{C,j} \leq 0.25$.

3.2.3. Décision

Chaque couple (A, I) de lignes de la matrice X se voit donc affecter par cet opérateur un gène C. Ce couple est considéré comme un couple activateur/inhibiteur d'un gène de X s'il existe dans cette matrice une ligne proche, au sens d'une mesure de similarité S, de C : plus précisément, nous proposons de normaliser C, et de retenir comme gènes candidats les gènes R maximisant :

$$S = 1 - \sum_{j=1}^{p} \left(x'_{Cj} - x_{Rj} \right)^2 / 4p. \qquad (6)$$

où $x'_{ij} = (x_{ij}-m_C)/ s_C$, m_C (resp.s_C) étant la moyenne (resp. la variance) de la ligne C.

4. Résultats et discussions

4.1. Analyse en composantes principales fonctionnelles

4.1.1. Exemple illustratif

Nous souhaitons dans un premier temps illustrer la méthode, et nous nous inspirons pour cela d'un exemple proposé par Han [HAN99], qui applique le même principe d'analyse à l'étude de la déformation spatio-temporelle du myocarde. Les fonctions $\phi_k(s)$ de base choisies sont des B-splines cubiques uniformes, performantes en terme d'interpolation et en rapidité de calcul.

Soient X_i, $1 \leq i \leq 5$, cinq courbes sinusoïdales, superposées à la première bissectrice $x(s) = s$, s appartenant à [0..19], et perturbées par une bruit gaussien additif, modélisé par une variable aléatoire g suivant une loi $N(0,0.5)$ (Figure 2-24) :

$(\forall s \in [0..19])$ $X_1 : x_1(s) = 4\sin(2\pi s/19) - s + g$

$X_2 : x_2(s) = -4\sin(2\pi s/19) - s + g$

$X_3 : x_3(s) = 2\sin(4\pi s/19) - s + g$

$X_4 : x_4(s) = -2\sin(4\pi s/19) - s + g$

$X_5 : x_5(s) = \sin(2\pi s/19) - s + g$

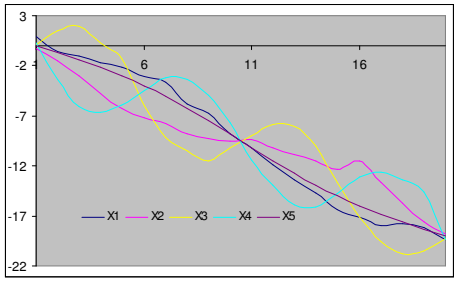
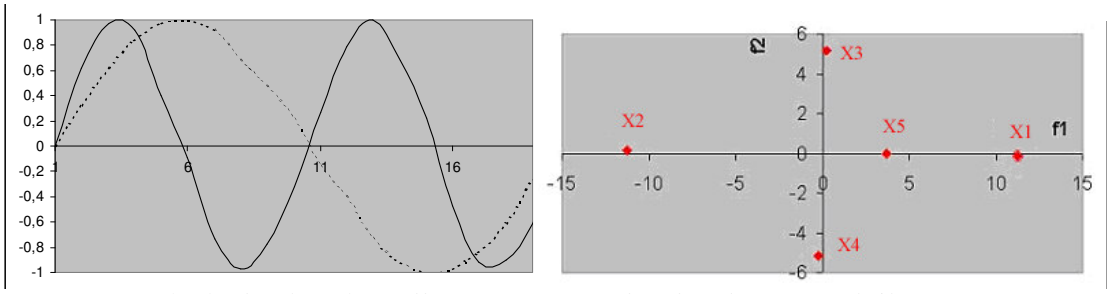


Figure 2-24 : exemple d'illustration de l'ACPF

Un échantillonnage a été réalisé dans une matrice 5×20, pour se rapprocher des données que nous traitons dans ce mémoire. Les courbes $x_1(s)$, $x_3(s)$ et $x_5(s)$ ont été échantillonnées par pas de 1 sur [0,19], les courbes $x_2(s)$ et $x_4(s)$ ayant été représentées avec 5 valeurs manquantes (des 0), choisies aléatoirement dans [0..19], modélisant par exemple une sous expression de gène ou une non détection du niveau de fluorescence.

Les deux premières composantes principales calculées contribuent à 99.9% de l'explication des variations des données (Figure 2-25-a). La première (resp. seconde) exprime une variation sinusoïdale de période 20 (resp. 10). Les projections des courbes X_i sur le plan de ces deux composantes (Figure 2-25-b) fait apparaître que les points X_3 et X_4 sont en opposition sur l'axe f_2 à une distance d_1=5.6 de l'origine. De même, les points représentant les courbes $x_1(s)$ et $x_2(s)$ sont en opposition sur l'axe f_1 à une distance $d_2 \approx 11.5$ de l'origine. Les mesures d_1 et d_2 expriment l'amplitude de la variation d'une courbe donnée par rapport à la courbe moyenne, d_2 étant deux fois plus grand que d_1, reflétant le rapport des amplitudes des courbes correspondantes. De plus, la distance du point X_5 à l'origine est $d \approx 3$, exprimant le rapport de ¼ des amplitudes de $x_5(s)$ et $x_1(s)$. Enfin, puisque X_3 varie de la même manière que la seconde composante principale, son score f_2 est positif (et le score f_2 de X_4 est négatif, puisque $x_4(s)$ est en opposition de phase avec la seconde composante principale fonctionnelle).



a- Composantes principales fonctionnelles b- Plan des scores f1/f2

Figure 2-25 : résultats de l'ACPF sur les données d'exemple

Notons enfin que les données manquantes n'ont eu aucune influence sur la classification des courbes finales, un des intérêts de l'ACPF résidant dans l'étape d'interpolation par les fonctions $\phi_k(s)$.

4.1.2. Données réelles

L'étude prospective de l'analyse en composantes principales fonctionnelles sur des données réelles a été effectuée sur deux jeux de données publiques : l'une concerne les données de sporulation de la levure *Saccharomyces cerevisiae*, l'autre des données extraites de 60 lignées de cellules tumorales.

- Sporulation de la levure

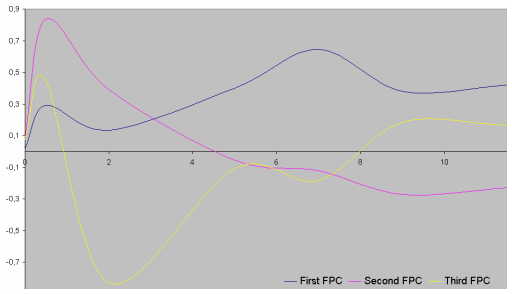
Les données sont constituées des ratios d'expressions de 6118 gènes connus où à la fonction présumée, mesurées à sept instants différents (0min, 30min, 2 h, 5h, 7h, 9h, 11.5h), de la levure Saccharomyces cerevisiae Ce champignon de la famille des Ascomycètes a la particularité d'être à la fois eucaryote et unicellulaire, ce qui en fait un outil de choix pour l'étude simplifiée des eucaryotes. Cette levure s'est de plus imposée comme modèle parce qu'elle est non pathogène, aérobie facultative, son temps de génération est court (environ 2 heures), et elle est bien adaptée à l'étude des croisements génétiques. Chu et al. [CHU98] ont identifié sept motifs temporels de transcription signant la progression de la sporulation.

Selon ces mêmes auteurs, nous n'avons retenu que les données des gènes présentant des variations significatives, i.e. ceux dont l'amplitude de variation est au moins de 2.2. Sur les 1266 gènes restants, environ 50% sont induits durant la sporulation.

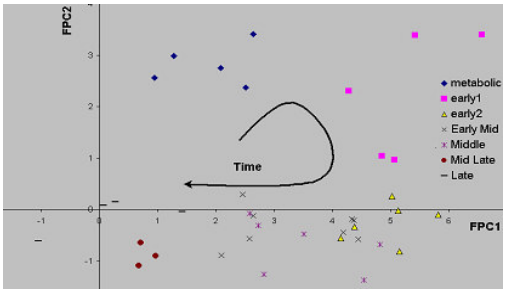
Les données se présentent donc sous la forme d'une matrice 1266×7 que nous analysons par ACPF. Les trois premières composantes fonctionnelles capturent 97% de la variance totale des données (Figure 2-26-a). Les gènes ayant un score élevé (resp. faible) sur la première composante sont fortement régulés positivement (resp. négativement) durant la sporulation. La seconde composante met en exergue un ordre dans la vitesse d'induction : plus l'expression du gène est induite tôt, plus le score sur cette composante s'est révélé faible. Enfin, la troisième composante évalue la vitesse des changements de variation des ratios d'expressions des gènes au cours de l'expérience, le score sur cet axe étant d'autant plus faible que l'expression varie dans le temps.

La projection des 7 profils de gènes représentatifs trouvés par Chu et al. dans l'espace 3D des composantes principales fonctionnelles révèle une organisation spatiale de ces gènes (Figure 2-26-b à –d). Par exemple, le plan des scores 1/2 montre un arrangement des données dans le sens trigonométrique, suivant le temps d'induction, depuis l'état métabolique (f_{i1} faible et f_{i2} fort) jusqu'à l'état tardif (f_{i1} et f_{i2} faibles).

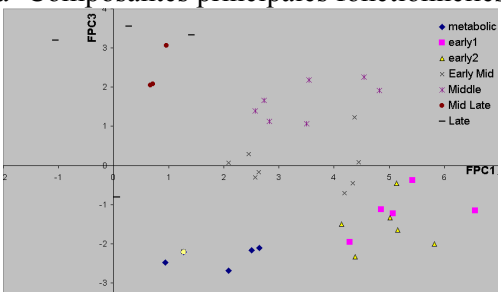
Les projections des 1266 gènes dans cet espace fonctionnel n'a pas révélé d'organisation précise en nuages de points, et comme Raychaudhuri *et al* [Raychaudh00], nous n'avons pas tenté d'appliquer un algorithme de classification dans cet espace de paramètres, préférant rechercher des correspondances fonctionnelles entre gènes et composantes principales, afin par exemple de prédire la fonction de gènes.



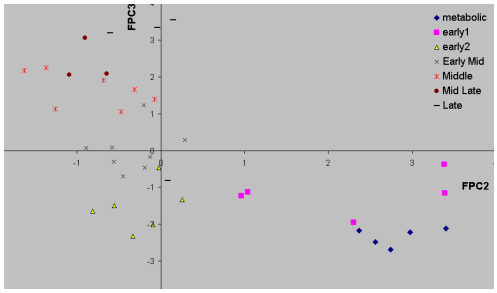
a- Composantes principales fonctionnelles



b- Plan des scores 1/2



c- Plan des scores 1/3



d- Plan des scores 2/3

Figure 2-26 : analyse des données de sporulation de la levure du boulanger par ACPF

- Lignées de cellules tumorales

Les données consistent ici en ARN messagers extraits de 60 lignées de cellules tumorales, et hybridées sur des puces à ADN contenant 9703 clones humains d'ADNc [ROSS00]. Ces derniers incluent environ 8000 gènes différents, dont 80% ont été identifiés. Nous analysons ici les données sous la forme d'une matrice 60×1416, incluant 1375 gènes distincts (filtrés avec un critère de variance maximale) et des répétitions. Les lignes de la matrice sont donc les lignées tumorales, sujettes aux variations induites par les gènes en colonnes. Le grand nombre de ces dernières interdit toute analyse visuelle immédiate et l'ACPF doit ici permettre à la fois de réduire la dimension de l'espace d'étude et d'exhiber des paramètres de variation spécifiques des lignées de cellules tumorales.

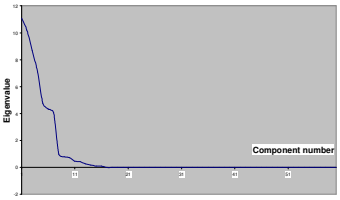
Le graphique de la Figure 2-27-a suggère la représentation des informations dans un espace à 6 dimensions, dont les composantes principales fonctionnelles associées sont représentées sur la Figure 2-27-b. La projection des courbes dans les plans de score permet d'identifier six groupes compacts de courbes (de cardinalités supérieures à 5), détaillés dans le Tableau 2-5, très proches d'un axe principal fonctionnel avec un score correspondant très élevé. Ces nuages de points regroupent des lignées de cellules dont les tissus d'origine présumés sont communs, avec quelques exceptions : les 7 lignées de tumeurs du colon ont été regroupées (avec NCI-H322M, associé au cancer du poumon) ; 7 lignées de cancer du rein (resp. lignées de mélanomes) ont été regroupées ; les 6 lignées de leucémie ont été rassemblées (avec une lignées T-47D du cancer du sein) ; 5 lignées du cancer des ovaires sur les 6 disponibles ont été regroupées (SK-OV3 a été mis dans un groupe contenant des lignées de cancer du sein), et toutes les lignées concernant le système nerveux central ont été rassemblées, avec deux lignées du cancer du sein. Les autres types de tumeur sont dispersés dans l'espace des scores, et sont reliés à des composantes mineures correspondant probablement à des réseaux d'activation d'un petit nombre de gènes.

Les composantes principales fonctionnelles éclairent quant aux variations principales des lignées de cellules tumorales qu'elles représentent le mieux. La lignée rénale SN-12C ayant en particulier un profil très différent de la composante principale agrégeant les autres tumeurs rénales (Figure 2-27-c) ne pouvait en particulier être classée dans le même nuage de points. Ce résultat est d'ailleurs retrouvé chez Ross et al. [ROSS00] et Crescenzi *et al*. [CRESCENZI01]. L'inclusion de NCI-H322M dans le nuage des lignées de cellules du colon provient du score très élevé de cette lignée sur la première composante principale fonctionnelle, celle-là même qui représente le mieux les données du colon (la distance de Malahanobis du

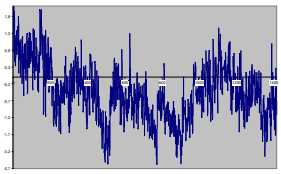
profil de NCI-H322M aux courbes du nuage est inférieure à 0.12, pour une distance de Malahanobis moyenne intra classe de 3.47). Les résultats obtenus sont en accord avec ceux déjà publiés, notamment par Ross et al [ROSS00] qui ont étudiés avec un algorithme de classification hiérarchique et une matrice de visualisation en pseudo-couleurs deux sous-ensembles de gènes (d'une part les 1167 gènes ayant les variations les plus importantes, d'autre part 6831 gènes bien identifiés), pour assurer la robustesse à leur analyse.

CO:KM12;CO:HCT-116; CO:SW620; CO:HCT15; LC:NCI-H322M;CO:HCC-2998; CO:HT29; CO:COLO205	LE:CCRF-CEM;LE:HL-60;LE:K-562;LE:RPMI-8226;LE:SR;LE-MOLT-4;BR:T-47D	ME:UACC-257;ME:SK-MEL-28;ME:MALME-3M;ME:SK-MEL-2; BR:MDA-MB-435; ME:M14;ME:UACC-62;ME:SK-MEL-5
RE:RXF-393;RE:TK-10;RE:ACHN;RE:A498;RE:786-0;RE:U0-31;RE:CAKI-1	OV-IGROV1; LC:NCI-H23; LC:NCI-H522; OV:OVCAR-5; OV:OVCAR-4; OV:OVCAR-3; OV:PC-3	CNS:SNB-75; CNS:SNB-19; BR-HS578T; CNS-U251; BR-BT-549; CNS:SF-268; CNS:SF-295; CNS:SF-539;

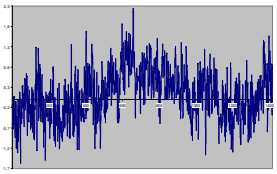
Tableau 2-5 : groupes de lignées tumorales détectés dans les plans de score



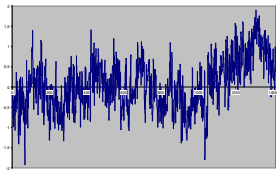
a- Graphique valeur propre (numéro de la composante)



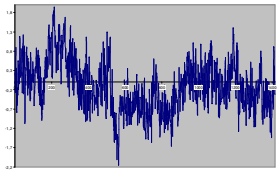
1ère composante (15%)



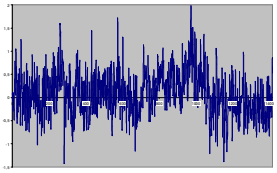
2ème composante (13%)



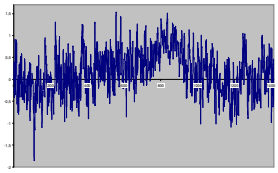
3ème composante (10%)



4ème composante (9%)

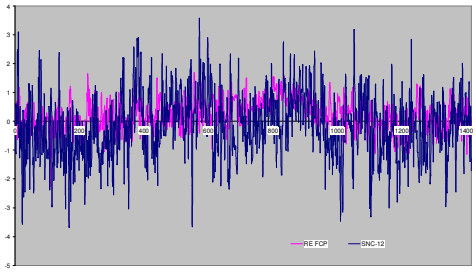


5ème composante (7%)



6ème composante (7%)

b- Composantes principales fonctionnelles (pourcentage de la variance expliquée)



c- Composante principale associée au rein et SNC-12

Figure 2-27 : analyse des données des lignées tumorales par ACPF

4.1.3. Discussion

L'intérêt de disposer de l'ACPF (par rapport à l'ACP classique par exemple) est que les données temporelles ne doivent pas nécessairement être échantillonnées aux mêmes instants : l'interpolation B-

spline permet de trouver, pour chaque temps, une valeur correspondante pour chaque composante principale fonctionnelle.
Au contraire des algorithmes classiques d'extraction de données utilisées en routine (classification ou discrimination), l'analyse en composantes principales fonctionnelles est une méthode privilégiant l'extraction de modes de variations significatifs des données, pouvant alors être corrélés à des processus biologiques particuliers en étudiant la corrélation entre les variables originales et les composantes principales fonctionnelles dans les plans de score. Ce qui rend l'ACPF intéressante est sa capacité de gérer des profils incomplets (correspondant à des données d'expression manquantes ou mal acquises), ou irrégulièrement acquis, ce point étant fondamental pour l'étude de données temporelles (*e.g.* cycle cellulaire). Pour ces mêmes études, l'ACPF est insensible aux données incomplètes (interpolation par les fonctions de base $\phi_k(s)$), contrairement à l'ACP classique dont l'application est de ce fait controversée dans ces situations [RAYCHAUDH00]. Notons que si certains auteurs proposent de combiner l'ACP à un algorithme supervisé de type Expectation-Maximization pour pallier les carences de l'ACP seule [TIPPING99], cette technique est en pratique lourde à mettre en œuvre.
Un autre intérêt de l'ACPF et de permettre une représentation analytique des composantes fonctionnelles, par l'intermédiaire de leur décomposition dans la base (ϕ_k). Chaque courbe, modélisant un profil d'expression de gène ou une composante principale, peut être analysée en utilisant des méthodes continues (de type analyse harmonique, si l'échantillonnage en expériences est suffisant), ceci par exemple pour extraire des paramètres caractéristiques des données d'entrée (e.g. fréquences dans le cas du cycle cellulaire).

4.2. Application de la fusion de données

4.2.1. Données simulées

Nous avons construit des matrices X simulant les données d'expression de N gènes sous p expériences. A partir de $0<n_A<25$ activateurs et $0<n_I<25$ inhibiteurs générés aléatoirement, placés sur les premières lignes de la matrice, nous avons calculé les expressions de n_D gènes dépendants par :

$$(\forall k \in [n_A+n_I+1,n_D])\ (\forall j \in [1,p])\ x_{k,j} = F((x_{a,j})_{1\le a\le nA}, (x_{i,j})_{nA+1\le i\le nA+nI}) \qquad (7)$$

où F est une fonction croissante en les $(x_{a,j}), 1\le a\le n_A$, et décroissante en les $(x_{i,j})$, $n_A+1\le i\le n_A+n_I$. Plus précisément, nous avons choisi :

$$F = 0 \sum_{A\subset[1..n_A]} \alpha_A . x_A - \sum_{I\subset[1..n_I]} \alpha_I . x_I \qquad (8)$$

Où A (resp. I) est un sous-ensemble d'indices tirés aléatoirement dans $[1..n_A]$ (resp. $[1..n_I]$), et les coefficients α sont tirés selon une loi normale $N(3,3)$, puis normalisés pour que leur somme soit égale à 1. Les lignes restantes de la matrice X ont été générées aléatoirement, et les lignes de la matrice ont été normalisées. L'ensemble des relations activateur/inhibiteur/gène simulées avec n_A=1, n_I=1, n_D=50, n=200, p=10 correspond aux plus grandes valeurs de similarité S calculées par l'algorithme. La Figure 2-28 donne un exemple du triplet (A, I, C) ayant obtenu la plus grande similarité (S=0.92) et le gène R résultat.

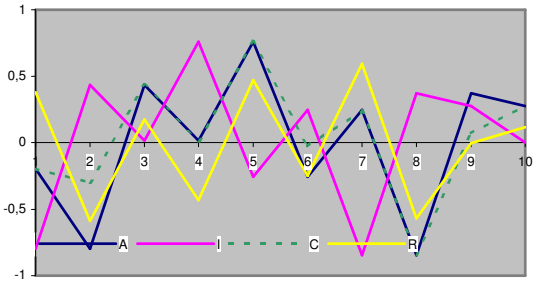


Figure 2-28 : exemple de profils d'un triplet (A, I, C)

L'algorithme a parfois échoué dans le cas où nous avons simulé simultanément des expressions fortes ou très fortes de l'activateur et de l'inhibiteur. Dans ce cas en effet, le gène résultant doit avoir un

comportement "moyen", que nous avons approché par une division du minimum des expressions observées sur *A* et *I*. L'expression résultante pour *C* n'est cependant pas nécessairement proche de cette valeur, et fait diminuer de façon critique la mesure de *S* (en particulier, une valeur "moyenne" peut être négative, alors que le calcul du minimum divisé donne toujours un réel positif). Sur l'ensemble des autres combinaisons possibles de comportements pour *A* et *I*, l'opérateur a toujours permis de décider sur le "bon gène", *i.e.* celui correspondant effectivement à la combinaison linéaire simulée.

Les techniques couramment utilisées dans l'analyse d'expression des gènes, et décrites dans les paragraphes précédents, ne permettent pas facilement de mettre en exergue une relation de type activateur/inhibiteur. Elles sont en effet conçues préférentiellement pour regrouper les gènes en familles ayant des comportements semblables, dans le but par exemple d'exhiber des groupes de gènes co-régulés ou participant à des fonctions biologiques semblables. A titre d'exemple, nous avons utilisé le logiciel Cluster [EISEN98] pour classer l'ensemble des données simulées dans un arbre phylogénique (classification hiérarchique, en utilisant comme mesure de similarité une corrélation, et un regroupement par la méthode du lien moyen). Le résultat est présenté sur la Figure 2-29, et montre que les gènes mis en jeu dans la relation sont très éloignés dans l'arbre résultat. Par contre, cet arbre permet de regrouper les gènes ayant des coefficients de combinaison linéaire proches et donc des profils similaires.

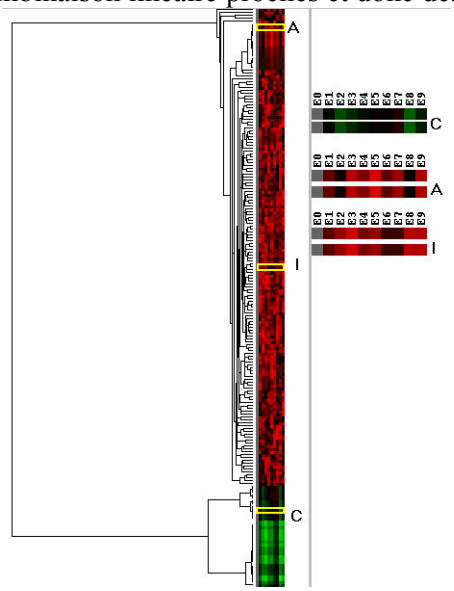


Figure 2-29 : classification hiérarchique sur les données de synthèse

Nous avons effectué des tests avec des jeux de valeurs différentes pour n_A et n_I et avons toujours retrouvé la majorité des relations liant activateurs/inhibiteurs et gènes résultants, même pour un nombre d'expériences faible (p=5). Dans le cas d'un nombre quelconque d'activateurs et d'inhibiteurs, l'opérateur *F* est un tableau à n_A+n_I entrées, dont la construction suit les règles intuitives du modèle activateurs/inhibiteurs. L'algorithme a cependant échoué sur les n-uplets comportant beaucoup de valeurs proches de 1, et la combinaison par une division du minimum devrait sans doute être reconsidérée dans ces situations. De plus, plus le nombre d'expériences p a été pris important, mieux les gènes ont été classés, cela étant sans doute dû au fait que l'augmentation de *p* induit une augmentation du nombre de variations dans les profils des gènes de *X*, et qu'il est donc plus difficile d'avoir deux gènes quelconques de la matrice ayant approximativement les mêmes profils. Seul le gène prototype *C* a alors un profil maximisant la mesure *S*. Notons enfin que dans la mesure où $n_A+n_I \leq p$, il est possible de calculer les coefficients de la combinaison linéaire décrivant au mieux (au sens de S) l'expression d'un gène en fonction de ses activateurs et de ses inhibiteurs.

4.2.2. Données réelles

Nous avons utilisé cet algorithme sur des données issues du séquençage du génome de la levure du boulanger *Saccharomyces cerevisiae*. Les données originales, disponibles sur http://cellcycle-

www.stanford.edu, décrivent l'expression relative d'ARN messager en fonction du temps dans des cultures cellulaires de levure synchronisées de quatre manières indépendantes : par le phéromone alpha (mesures toutes les 7 minutes pendant 119 minutes), par l'utilisation d'un mutant sensible à la température, cdc15, qui stoppe la méiose à la température critique (mesures toutes les 10 minutes pendant 290 minutes), par l'utilisation d'un mutant sensible à la température, cdc28 [CHO98] (mesures toutes les 10 minutes pendant 160 minutes) et par élutriation (mesures toutes les 30 minutes pendant 390 minutes). Quatre méthodes sont utilisées car chacune introduit des artefacts propres (*e.g.* choc thermique avec l'utilisation des mutants). Spellman *et al*. [SPELLMAN98] ont identifié 800 gènes régulés par le cycle cellulaire. Nous utilisons dans la suite 698 d'entre eux, pour lesquels aucune donnée manquante n'est observée sur les 72 conditions couvrant les quatre expériences (nous n'avons pas retenu, comme proposé par Tamayo *et al.* [TAMAYO99], l'expérience à 90 minutes du mutant cdc15), et nous normalisons cette matrice 698*72 de telle sorte que chaque ligne soit de moyenne nulle et de variance unité.
Là encore, nous avons cherché les gènes répondant à un modèle simple $n_A = n_I = 1$. Nous présentons sur la Figure 2-30 un exemple de gène ayant la plus grande similarité avec un des gènes prototype calculé, l'activateur étant le gène ROX3 et l'inhibiteur le gène HAP1.

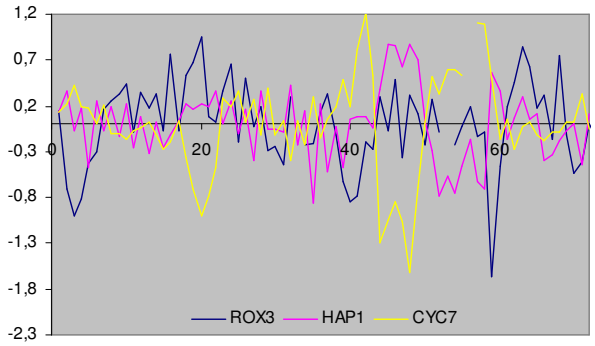


Figure 2-30 : exemple de triplet (A, I, C) obtenu sur le génome de *Saccharomyces cerevisiae*

HAP1 est un facteur de transcription codant pour une protéine qui joue un rôle important dans le contrôle de l'activation de la transcription par la pression d'oxygène, et qui se trouve ainsi impliquée dans la régulation de nombreux gènes codant pour des fonctions respiratoires. En particulier, HAP1 inhibe CYC7 dans des conditions anaérobies [PREZANT87]. ROX3 encode quant à lui une protéine impliquée dans l'expression du gène CYC7 [ROSENBLUM91]. D'après la littérature, ce triplet interagit bien comme prévu par l'algorithme, et ne pourrait être détecté par une méthode classique de regroupement, les profils des trois gènes étant bien différents.
Quarante sept triplets (A, I, C) ont été trouvés, pour lesquels le gène prototype C est très proche d'un gène de la matrice G (i.e. au moins 90% du maximum de similarité dans les données). Parmi ceux-ci, nous avons retrouvé un triplet (CYB2, HAP1, CYC7), déjà décrit dans [WOOLF00], et relié au processus de respiration de la levure. Nous sommes actuellement en cours de validation des autres triplets, en recherchant les fonctions biologiques des activateurs, inhibiteurs et gènes résultats correspondants.

4.2.3. Discussion

L'utilisation de la logique floue dans l'interprétation des données issues des puces à ADN permet de prendre en compte les ambiguïtés des mesures (imprécisions dues au bruit lors de la mesure, imprécision lors de la détection des spots sur la plaque…). L'utilisation d'une technique de combinaison inspirée de la fusion de données permet quant à elle à la fois de modéliser le comportement intuitif d'un modèle activateurs / inhibiteurs simple, et de retarder la décision prise sur C en autorisant l'apport d'informations supplémentaires (données biologiques, autres informations sur les activateurs/inhibiteurs), le but étant alors d'introduire le maximum de compétences pour augmenter la certitude du résultat sur C. Le modèle construit ici est simple et intuitif, et suppose que l'expression d'un gène dépend linéairement d'un ensemble d'activateurs et d'inhibiteurs. La réalité biologique est évidemment bien plus complexe, et

certains gènes peuvent par exemple interagir avec d'autres par l'intermédiaire de multiples sites de liaison, ce que l'algorithme ne pourra détecter. Néanmoins, cette première approche permet de capturer des relations entre gènes et de dresser un tableau fonctionnel simple d'interactions.
Pour un modèle impliquant un seul activateur et un unique inhibiteur, toutes les combinaisons de 2 gènes de X sont testées et une comparaison du résultat C à tous les gènes de X est effectuée. L'algorithme est donc en $O(N^3)$, ce qui peut être rédhibitoire en temps de calcul pour un nombre de gènes important, comme on trouve fréquemment sur les puces à ADN. De plus, si on veut étendre le modèle à un nombre quelconque d'activateurs et d'inhibiteurs, la combinatoire explose et il est en outre judicieux de s'interroger sur la validité de ce modèle pour un grand nombre d'interactions.
La meilleure compréhension des mécanismes de régulation des gènes a de nombreuses applications, comme par exemple la validation de cibles thérapeutiques dans le traitement d'une pathologie ou encore l'identification de protéines au rôle encore inconnu dans la régulation d'un processus biologique donné. Pour ce dernier cas par exemple, si l'algorithme détecte une régulation d'un gène R par un activateur A connu et un inhibiteur I encore méconnu dans le processus, alors des hypothèses pourront être formulées quant à la protéine encodée par I.

Conclusion

La richesse des données proposée par les puces à ADN nécessite le développement de méthodes d'extraction de connaissances adaptées, et d'outils de prétraitement ne conservant que l'information la plus pertinente. Ce chapitre a présenté les méthodes les plus couramment utilisées dans ce domaine, et a développé deux nouvelles techniques d'analyse des données extraites des images de puce à ADN. Chacune de ces méthodes a été validée sur des données de synthèse, puis appliquée sur des données réelles en fonction de ses champs d'applications propres. Les résultats obtenus sont en accord avec les résultats par ailleurs publiés sur les mêmes données.
Les futurs développements dans ce domaine au laboratoire incluent maintenant l'application de ces nouvelles méthodes sur les jeux de données disponibles localement, concernant l'étude du pathogène *Encephalitozoon cuniculi*. Et, poursuivant l'objectif du projet d'étude présenté dans l'introduction de cette seconde partie, l'intégration de ces algorithmes à un nouvel ensemble logiciel que nous développons au laboratoire.

Conclusion

Le projet de recherche décrit dans cette partie se propose donc de créer une plateforme de simulation d'images de puces à ADN, permettant aux bioinformaticiens de tester ou de comparer des méthodes d'analyses existantes ou à venir, et d'intégrer leurs propres outils. Développée au sein du LIMOS, cette activité propose tout d'abord la réalisation d'un simulateur d'images de puces à ADN, puis l'élaboration pour les diverses phases de traitement de ces images de nouveaux outils d'analyse. Quelques résultats ont d'ores et déjà été obtenus, et concernent d'une part la recherche des limites des méthodes existantes, et d'autre part l'intégration de nouveaux algorithmes de segmentation et d'extraction de connaissances validées.

Les perspectives de travail dans ce domaine sont nombreuses et variées. Il paraît tout d'abord indispensable de diffuser au plus grand nombre le simulateur proposé dans le chapitre 1, *via* par exemple la plateforme WEB mise en place au laboratoire, ceci pour valider la modélisation effectuée et proposer d'éventuelles améliorations et la prise en compte d'autres informations (par exemple type de support). Cette phase de validation multicentrique autorisera ensuite à considérer les images de biopuces générées comme suffisamment réalistes pour évaluer de façon pertinente les méthodes existantes et les algorithmes émergents. Les tests se devront d'être systématiques, et porteront sur les grandes étapes de pré traitement, de détection, de quantification et de classification des données. Le jeu de paramètres des simulations permettra non seulement d'évaluer les différents outils disponibles dans des conditions variées, mais aussi de proposer des domaines de validité pour les différentes méthodes utilisées. Une collaboration avec les différents laboratoires possédant les logiciels est ici envisagée.

Plus généralement, nous envisageons dans cette phase non seulement de tester les logiciels d'analyse, mais aussi de reprogrammer les méthodes classiques (*e.g.* localisation des spots par cercle fixe ou variable, analyse d'histogramme, algorithmes de classification des données) et de modifier leur structure pour les adapter aux jeux de paramètres des simulations, ceci par exemple dans le but de proposer des algorithmes adaptatifs en fonction des conditions de l'expérience.

Enfin, à la lumière des analyses de la phase précédente, il est prévu à plus long terme de proposer un nouvel ensemble logiciel réalisant l'analyse complète des images de puces et intégrant au mieux des réponses aux différents problèmes soulevés précédemment. Cet outil pourra se concrétiser sous deux versions, l'une logicielle et l'autre en ligne, cette dernière pouvant par exemple être intégrée au site WEB réalisé dans la phase 1. Il est également envisagé dans cette dernière étape d'associer à chaque collection de gènes déduits de la classification l'ensemble des propriétés retirées des banques de données publiques accessibles sur Internet. L'objectif étant alors par exemple de déduire des fonctions biochimiques pour cet ensemble.

BIBLIOGRAPHIE

[ABE97] ABE K., FUJIMURA H., TOYOOKA K., SAKODA S., YORIFUJI S., YANAGIHARA T., Cognitive function in amyotrophic lateral sclerosis, *Journal of Neurological Sciences*, **148**:95-100, 1997.

[ALIZADEH00] ALIZADEH A., EISEN M., DAVIS R., MA C., LOSSOS I., ROSENWALD A., BOLDRICK J., SABET H., TRAN T., YU X., POWELL J., YANG L., MARTI G., MOORE T., HUDSON J., LU L., LEWIS D., TIBSHIRANI R., SHERLOCK G., CHAN W., GREINER T., WEISENBURGER D., ARMITAGE J., WARNKE R., STAUDT L., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**:503-511, 2000.

[ALTER03] ALTER O., O'BROWN P., BOTSTEIN D., Singular value decomposition for genome-wide expression data processing and modelling, *Proceedings of the National Academy of Sciences*, **97**:10101-10106, 2000.

[ANGULO03] ANGULO J., SERRA J., Automatic analysis of DNA microarray images using mathematical morphology, *Bioinformatics*, **19**:553-562, 2003.

[AUSSEM02] AUSSEM A., PETIT JM., ε-functional dependency inference: application to DNA microarray expression data, *Proceedings of BDA'02*, 2002.

[AZUAJE01] AZUAJE F., A computational neural approach to support the discovery of gene function and classes of cancer, *IEEE Transactions on Biomedical Engineering*, **48**:332-339, 2001.

[BAJCSY89] BAJCSY B., Multiresolution elastic matching, *Computer Vision, Graphics and Image Processing*, **46**:1-21, 1989.

[BALAGURUN02] BALAGURUNATHAN Y, DOUGHERTY E, CHEN Y, BITTNER MJ, TRENT JM., Simulation of cDNA microarrays via a parameterized random signal model, *Journal of Biomedical Optics*, **7**: 507-523, 2002.

[BARASH02] BARASH Y., FRIEDMAN N., Context-Specific Bayesian Clustering for Gene Expression Data, *Journal of Computational Biology*, **9**:169-191, 2002.

[BARKER85] BARKER A., FREESTON I., JALINOUS R., MERTON P., MORTON H., Magnetic stimulation of the human brain, *Journal of Physiology*, **369**:3P, 1985

[BARRA96] BARRA V, MORIO B, COLIN A, VERMOREL M, BOIRE JY, Automatic Assessment of Muscle/Fat temporal Variations on MR Images of the Thigh, *Proceedings of IEEE Engineering in Medicine and Biology Society, 18th Annual International Conference*, Amsterdam, pp. 247-248, 1996.

[BARRA98-A] BARRA V., COLIN A., BOIRE J.Y., Synthesis of a High Resolution functional Image by an MR/SPECT Fusion Process, *European Journal of Nuclear Medicine, Proceedings of the European Association of Nuclear Medicine Annual Congress*, Berlin, 490, 1998.

[BARRA98-B] BARRA V., BOIRE J.Y., Caractérisation des Tissus Cérébraux en IRM par Classification Possibiliste de Propriétés de Voxels, *Actes de la sixième Rencontre de la Société Francophone de Classification,* Montpellier, 15-19, 1998.

[BARRA98-C] BARRA V, DATIN C, COLIN A, BONNY JM, LAURENT W, SARRY L, ATTAK M, BOIRE JY, RENOU JP, Segmentation automatique des Composantes structurales de la Viande à partir d'Images RMN, *Actes des dixièmes rencontres AGORAL*, Massy, pp. 265-270, 1998.

[BARRA99] BARRA V., BOIRE J.Y., Segmentation floue des tissus cérébraux en IRM 3D. Approche possibiliste *versus* autres méthodes, *Rencontres Francophones sur la Logique Floue et ses Applications*, Valenciennes, Editions Cépaduès, 193-198, 1999.

[BARRA00-A] BARRA V., Fusion d'images 3D du cerveau : étude de modèles et applications, *thèse, Universités de Clermont-Ferrand*, 2000.

[BARRA00-B] BARRA V., BOIRE J.Y., Tissue Segmentation on MR Images by a possibilistic Clustering on a 3D Wavelet Representation, *Journal of Magnetic Resonance Imaging*, **11**: 267-278, 2000.

[BARRA00-C] BARRA V., BOIRE J.Y., MR Images Fusion for Brain Tissues Volume Measurement, *Proceedings of the 3rd International Conference on Computer Vision, Pattern Recognition and Image Processing*, Atlantic City, pp. 362-366, 2000.

[BARRA00-D] BARRA V., BOIRE J.Y., Quantification of brain tissue Volumes using MR/MR Fusion, *Medical Physics, Proceedings of the World Congress on Medical Physics and Biomedical Engineering,* Chicago, 1404-1410, 2000.

[BARRA00-E] BARRA V., BOIRE J.Y., Evaluation d'Opérateurs de Fusion en Imagerie médicale : exemple de la Fusion d'Images IRM, *Rencontres Francophones sur la Logique Floue et ses Applications*, La Rochelle, Editions Cépaduès, 139-145, 2000.

[BARRA00-F] BARRA V., BOIRE J.Y., Correction of Partial Volume Effects in SPECT using a MR/SPECT Fusion Process, *European Association of Nuclear Medicine Annual Congress*, *Paris*, 2-6 Septembre 2000.

[BARRA00-G] BARRA V., BOIRE J.Y., Aggregation of anatomical and functional Information by a MR/SPECT Fusion Process: Application to neurodegenerative Pathologies, *Proceedings of the Sixth Annual Meeting of the Organization for human Brain Mapping*, San Antonio, p. 633, 2000.

[BARRA00-H] BARRA V., BOIRE J.Y., Une approche possibiliste de la Fusion d'Images médicales, *Actes du douzième Congrès Francophone AFRIR-AFIA, RFIA 2000*, Paris, 309-316, 2000

[BARRA01-A] BARRA V., BOIRE J.Y., Automatic Segmentation of subcortical brain Structures in MR Images using Information Fusion, *IEEE Transactions on Medical Imaging*, **20**: 549-58, 2001.

[BARRA01-B] BARRA V., LEMAIRE J.J., DURIF F., BOIRE J.Y., Segmentation of the subthalamic Nucleus in MR Images using Information Fusion: a preliminary Study for a computed-aided Surgery of Parkinson's Disease, *Lecture Notes in Computer Science, Proceedings of MICCAI 2001*, Utrecht, 1183-1184, 2001.

[BARRA01-C] BARRA V., BRIANDET P., BOIRE J.Y., Fusion in medical Imaging: Theory, Interests and industrial Applications, *MedInfo,* **10**: 896-900, 2001.

[BARRA01-D] BARRA V., BOIRE J.Y., A General Framework for the Fusion of Anatomical and Functional Medical Images, *NeuroImage*, **13** : 410-424, 2001.

[BARRA02-A] BARRA V., Application de la fusion de données à l'analyse de données d'expression de gènes, *Rencontres Francophones sur la Logique Floue et ses Applications, Montpellier*, Editions Cépaduès, pp. 147-154, 2002.

[BARRA02-B] BARRA V., FRENOUX E., BOIRE J.Y., Automatic volumetric measurement of lateral ventricles on MR images with correction of partial volume effects, *Journal of Magnetic Resonance Imaging*, **15**: 16-22, 2002.

[BARRA02-C] BARRA V, BOIRE JY, Segmentation of Fat and Muscle from MR Images of the Thigh by a possibilistic clustering Algorithm, *Computer Methods and Programs in Biomedicine* **68**:185-193, 2002.

[BARRA03] BARRA V., GOUINAUD C., Simulation of Microarray Experiments, *Summer Computer Simulation Conference*, Montreal, pp 115-118, 2003

[BARRA04-A] BARRA V., Analysis of Gene Expression Data using Functional Principal *Components, Computer Methods and Programs in Biomedicine*, 2004. (à paraître)

[BARRA04-B] BARRA V., GOUINAUD C., KERANDEL A., Simulation of microarray image experiments – from theory to the Web-based application, *Simulation* (soumis)

[BARRA04-C] BARRA V., Segmentation of microarray images using a split and merge algorithm based on Delaunay triangulation, *Bioinformatics*, soumis.

[BASARSKY00] BASARSKY T., VERDNIK D., ZHAI J.YWELLIS D., Overview of a microarray scanner: Design essentials for an integrated acquisition and analysis platform, in Schena M (ed): *Microarray Boichip Technology,* Eaton Natick, 265-284, 2000.

[BENDOR99] BEN-DOR, A, SHAMIR, R, YAKHINI, Z. Clustering gene expression patterns, *Journal of Computational Biology*, **6**:281-297, 1999.

[BISHOP74] BISHOP JO, MORTON JG et al., Three abundance classes in Hela cell messenger RNA, *Nature*, **250**: 199-240, 1974.

[BLAND86] BLAND M., ALTMAN D., Statistical methods for assessing agreement between two methods of clinical measurement, *The Lancet*, **8**:307-10, 1986.

[BLOCH96] BLOCH I., Information Combination Operators for Data Fusion : A Comparative Review with Classification, *IEEE Transactions on Systems, Man, and Cybernetics*,**1**: 52-67, 1996.

[BLOCH03] BLOCH I., GERAUD T., MAITRE H., Representation and fusion of heterogeneous fuzzy information in the 3D space for model-based structural recognition-Application to 3D brain imaging, *Artificial Intelligence*, **148**: 141-175, 2003.

[BOUSSION03] BOUSSION N, CINOTTI L, BARRA V, RYVLIN P, MAUGUIERE F, Extraction of epileptogenic foci from PET and SPECT images by fuzzy modeling and data fusion, *NeuroImage,* **19**: 645-654, 2003.

[BOZINOV02] BOZINOV D, RAHNENFUHRER J., Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering, *Bioinformatics*, **18**: 747-756, 2002.

[BRÄNDLE00] BRÄNDLE N., CHEN Y., BISCHOF H., LAPP H., Robust Spot Fitting for Genetic Spot Array Images, *Proceedings of ICIP2000 - International Conference on Image Processing*, Vancouver, 412-415, 2000.

[BRANDT94] BRANDT M., BOHAN T., KRAMER L., FLETCHER J., Estimation of CSF, white and gray matter volumes in hydrocephalic children using fuzzy clustering of MR images, *Computer Medical Imaging and Graphics*, **18**:25-34, 1994.

[BRANSTON91] BRANSTON N., TOFTS P., Analysis of the distribution of currents induced by a changing magnetic field in a volume conductor, *Physics in Medicine and Biology*, **36**:161-168, 1991.

[BROWN97] BROWN M., GRUNDY W., LIN D., CRISTIANINI N., SUGNET C., FUREY T., ARES M., HAUSSLER D., Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proceedings of the National Academy of Sciences*, **97**: 262-267, 1997.

[CERRI95] CERRI G., DE LEO R., MOGLIE F., SCHIAVONI A., An accurate 3D model for magnetic stimulation of the brain cortex, *Journal of Medicine and Engineering Technology*, **19**:7-16, 1995.

[CHEN97] CHEN Y., DOUGHERTY E., BITTNER M., Ratio-based decision and the quantitative analysis of cDNA microarray images, *Journal of Biomedical Optics*, **2**: 364-374, 1997.

[CHENG00] CHENG Y., CHURCH G., Biclustering of expression data, *Proceedings of ISMB'2000*, AAAI Press, 93-103, 2000.

[CHO98] CHO R.J., CAMPBELL M.J., WINZELER E.A., STEINMETZ L., CONWAY A., WODICA L., WOLFSBERG T.G., GABRIELIAN A., LANDSMAN D., DAVIS R., A genomic-wide transcriptional analysis of the mitotic cell cycle, *Molecular Cell*, **2**: 65-73, 1998.

[CHU98] CHU S., DERISI J., EISEN M., MULHOLLAND J., BOTSTEIN D., BROWN P., HERKOWITZ I., The transcriptional program of sporulation in budding yeast, *Science*, **282**:699-705, 1998.

[COHEN89] COHEN L., ROTH B., NILSSON J., DANG N., PANIZZA M., BANDINELLI S., FRIAUF W., HALLETT M., Effects of coil design on delivery of focal magnetic stimulation. Technical considerations, *Electroencephalography and Clinical Neurophysiology*, **75**:350-357, 1989.

[COLIN97] COLIN A, Etude de Méthodes de Recalage et de Fusion d'Images 3D du Cerveau. Application au suivi d'une Pathologie Cérébrale, *Thèse, Université de Clermont-Ferrand,* 1997.

[COPE03] COPE L., IRIZARRY R., JAFFEE H., WU Z., SPEED T., A Benchmark for Affymetrix GeneChip Expression Measures, *Bioinformatics*, **20**: 304-314, 2003.

[CRESCENZI01] CRESCENZI M., GIULIANI M., The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data, *FEBS Letters*, **507**:114-118, 2001.

[DAVEY03] DAVEY K., EPSTEIN C., GEORGE M., BOHNING D., Modeling the effects of electrical conductivity of the head on the induced electric field in the brain during magnetic stimulation, *Clinical Neurophysiology*, **114**:2204-2209, 2003.

[DEMONGEOT02] DEMONGEOT J., FRANÇOISE J.P. RICHARD M., SENEGAS F. BAUMA T.P., A differential geometry approach for biomedical image processing, *Comptes Rendus Biologies ,* **325**: 367-374, 2002.

[DELLEPIANE92] DELLEPIANE S., VENTURI G., VERNAZZA G., Model generation and model matching of real images by a fuzzy approach, *Pattern Recognition*, **25**: 115-137, 1992.

[DERISI00] DERISI J, VAN DEN HAZEL B, MARC P, BALZI E, BROWN P, JACQ C, GOFFEAU A., Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants, *FEBS Letters*, **24**:156-160, 2000.

[DOKLADAL03] DOKLADAL P., BLOCH I., COUPRIE M., RUIJTERS D., USTASUN R., GARNERO L., Topologically controlled segmentation of 3D magnetic resonance images of the head by using morphological operators, *Pattern Recognition*, **36**:2463-2478, 2003.

[DORSEL99] DORSEL A., Fundamental performance limitations of hybridized arrays, Invited paper, *Lake Tahoe Symposium on microarray algorithms and statistical analysis: methods and standards*, 1999.

[DROR01] DROR R., Noise Models in Gene Array Analysis, *Internal report*, MIT Department of Electrical Engineering and Computer Science, 2001.

[DUBOIS88] DUBOIS D, PRADE H, Possibility Theory, an approach to the computerized processing of uncertainty, Plenum Press, 1988.

[DUDOIT02] DUDOIT S., FRIDLYAND J., SPEED T., Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**:77-87, 2002.

[DUTA98] DUTA N., SONKA M., Segmentation and interpretation of MR brain images: an improved active shape model, *IEEE Transactions on Medical Imaging*, **17**:1049-1062, 1998.

[EATON92] EATON H., Electric field induced in a spherical volume conductor from arbitrary coils: application to magnetic stimulation and MAG, *Medicine and Biology Engineering Computation*, **30**: 433-440, 1992.

[EISEN98] EISEN, M., SPELLMAN, P., BROWN, P., BOTSTEIN, D. Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences*, **95**: 14863-14868, 1998.

[ELFAKHRI01] EL FAKHRI G., MOORE S., MAKSUD P., AURENGO A., KIJEWSKI M., Absolute Activity Quantitation in Simultaneous $^{123}I/^{99m}Tc$ Brain SPECT, *The Journal of Nuclear Medicine,* **42**:300-308, 2001.

[ERGÜT03] ERGÜT E., YARDIMCI Y., MUMCUOĞLU E., KONU O., Analysis of microarray images using FCM and K-means clustering algorithms, *Proceedings of International Conference on Signal Processing*, **1-2**:025, 2003.

[ESSELLE95] ESSELLE K., STUCHLY M., Cylindrical tissue model for magnetic field stimulation of neurons: effects of coil geometry, *IEEE Transactions on Biomedical Engineering*, **42**:934-941, 1995.

[FILIPEK91] FILIPEK P., KENNEDY D., CAVINESS JR V., Volumetric analyses of central nervous system neoplasm based on MRI, *Pediatric Neurology*, **7**: 347-51, 1991.

[FRANSSEN02] FRANSSEN-VAN HAL NL, VORST O, KRAMER E, HALL RD, KEIJER J., Factors influencing cDNA microarray hybridization on silylated glass slides, *Anals of Biochemistry*, **308**:5-17, 2002.

[FRENOUX01] FRENOUX E., BARRA V., BOIRE J.Y., Segmentation of the striatum using data fusion, 23rd Annual *International Conference of the IEEE Engineering in Medicine and Biology Society,* Istanbul, 25-28, 2001.

[FRENOUX02-A] FRENOUX E., BARRA V., BOIRE J.Y., Segmentation du striatum par fusion d'informations numériques et symboliques, *in Informatique et Santé*, SPRINGER, 180-189, 2002.

[FRENOUX02-B] FRENOUX E., BARRA V., BOIRE J.Y., Quantification of neurotransmission defects in functional imaging using information fusion: a prospective study, *9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Annecy, 1595-1596, 2002.

[FRENOUX03-A] FRENOUX E., BARRA V., BOIRE J.Y., HABERT M.O, A new method for the quantitative study of neurotransmission, *25th IEEE Engeneering in Medicine and Biology Society*, Cancun, Mexique, septembre 2003.

[FRENOUX03-B] FRENOUX E., Applications de la fusion d'images à l'étude de pathologies cérébrales *Thèse de doctorat*, Université d'Auvergne, 2003.

[FRENOUX04] FRENOUX E., BARRA V., BOIRE J.Y., Automatic cerebellum segmentation on magnetic resonance images using data fusion, *Journal of Magnetic Resonance Imaging*, 2003 (soumis).

[FUREY00] FUREY TS, DUFFY N, CRISTIANINI N, BEDNARSKI D, SCHUMMER M, HAUSSLER D, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, *Bioinformatics*, **16**:906-914, 2000.

[GABRIEL01] GABRIEL C., GABRIEL S., Compilation of the Dielectric Properties of Body Tissues at RF and Microwave Frequencies, *Technical report AL/OE-TR-1996-0037*, Brooks Air Force Laboratory, 2001.

[GEDDES67] GEDDES L., BAKER L., The specific resistance of biological material – a compedium of data for the biomedical engineer and physiologist, *Medicine and biology engineering*, **5**:271-293, 1967.

[GERAUD98] GERAUD T., Segmentation des structures internes du cerveau en IRM, *thèse, Ecole Nationale Supérieure des Télécommunications de Paris*, 1998.

[GETZ00] GETZ G, LEVINE L, DOMANY E, Coupled two-ways clustering analysis of gene microarray data, *Proceedings of the National Academy of Sciences*, **97**: 12079-12084, 2000.

[GHANEI98] GHANEI A., SOLTANIAN-ZADEH H., WINDHAM J., Segmentation of the hippocampus from brain MRI using deformable contours, *Computer Medical Imaging and Graphics*, **22**: 203-216, 1998.

[GHOSH02] GHOSH D., CHINNAIYAN A., Mixture modelling of gene expression data from microarray experiments, *Bioinformatics*, **18**:275-286, 2002.

[GIBAUD97] GIBAUD B., GERLATTI S., BARRILOT C., FAURE E., Methodology for the design of digital brain atlases, *Artificial Intelligence in Medicine*, E. Keravnou *et al* editors, Grenoble, 441-451, 1997

[GOLUB99] GOLUB T., SLONIM D., TAMAYO T., HUARD C., GAASENBEEK M., MESIROV H., COLLER JP., LOH M., DOWNING J., CALIGIURI M., BLOOMFIELD C., LANDER E., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**:531-537, 1999.

[GOODMAN76] GOODMAN JW, Some fundamental properties of speckle, *Journal of the Optical Society of America*, **66**: 1145-1150, 1976.

[GRANDORI91] GRANDORI F., RAVAZZANI P., Magnetic stimulation of the motor cortex – theoretical considerations, *IEEE Transactions on Biomedical Engineering*, **38**:180-191, 1991.

[GUNNING98] GUNNING-DIXON F., HEAD D. MCQUAIN J., ACKER J., RAZ N., Differential aging of the human striatum: a prospective MR imaging study, *American Journal of Neuroradiology*, **19**:1501-1507, 1998.

[GUYON01] GUYON I, WESTON J, BARNHILL S, VAPNIK V, Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, **46**:389-422. 2001.

[GUTHKE01] GUTHKE R., SCHMIDT-HECK W., HAHN D., PFAFF M., Gene Expression Data Mining for Functional Genomics using fuzzy technology, in *International Series in Intelligent Technologies: Advances in Computational Intelligence and Learning: Methods and Applications*, Kluwer Academic Publishers, 2001.

[HAMPEL02] HAMPEL H., TEIPEL S., BAYER W., ALEXANDER G., SCHWARZ R., SCHAPIRO M., RAPOPORT S., MÖLLER H., Age transformation of combined hippocampus and amygdala volume improves diagnostic accuracy in Alzheimer's disease, *Journal of Neurological Science*, **194**:15-19, 2002.

[HAN99] HAN M., Analyse Exploratoire de la Déformation Spatio-temporelle du Myocarde à partir de l'Imagerie par Résonance Magnétique de Marquage Tissulaire, *Thèse de doctorat*, INSA Lyon, 1999.

[HANDRAN01] HANDRAN S., WANG C., AZIZ D., Assessing Slide Flatness, *Application Note*, Axon Instruments Inc, 2001.

[HARTMANN99] HARTMANN S., PARKS M., MARTIN P., DAWANT B., Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations: Part II, validation on severely atrophied brains, *IEEE Transactions on Medical Imaging*, **18**: 917-926, 1999.

[HARTUV99] HARTUV E., SCHMITT A., LANGE J., MEIER-EWERT S., LEHRACHN H., SHAMIR R., An algorithm for clustering cDNAs for gene expression analysis, *Proceedings of the third international conference on Computational Molecular biology*, RECOMB, 1999.

[HASHIMOTO95] HASHIMOTO M., OHTSUKA K., TMS over the posterior cerebellum during visually guided saccades in man, *Brain*, **118**: 1185-1193, 1995.

[HASTIE00] HASTIE T., TIBSHIRANI R., EISEN M., ALIZADEH A., LEVY R., STAUDT L., CHAN W., BOTSTEIN D., BROWN P., Gene shaving as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology*, **1**:3-21, 2000.

[HAUEISEN95] HAUEISEN J., RAMON C., EISELT M., BRAUER H., NOWAK H., Influence of tissue resistivities on neuromagnetic fields and potentials studied with a finite element and a boundary element method of the head, *IEEE Transactions on biomedical engineering*, **23**, 1995.

[HEINZ92] HEINZ HÖHNE K., HANSON W., Interactive 3D segmentation of MRI and CT volumes usong morphological operations, *Journal of Computer Assisted Tomography*, **16**:285-294, 1992.

[HELD97] HELD K., ROTA H., KRAUSE B., WELLS W., KIKINIS R., MULLER-GÄRTNER H., Markov random field segmentation of brain MR images, *IEEE Transactions on Medical Imaging*, **16**:878-886, 1997.

[HERRERO01] HERRERO J., VALENCIA A., DOPAZO J., A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics*, **17**:126-136, 2001.

[HERWIG99] HERWIG, R, POUSTKA, AJ, MÜLLER, C, BULL, C, LEHRACH, H, O'BRIEN, J. Large-scale clustering of cDNA-fingerprinting data. *Genome Research*, **9**:1093-1105, 1999.

[HESSNER03] HESSNER M.J., WANG X., KHAN S., MEYER L., SCHLICHT M., TACKES J., DATTA M.W., JACOB H.J., GHOSH S., Use of a three-color cDNA microarray platform to measure and control support-bound probe for improved data quality and reproducibility, *Nucleic Acids Research* **31**:e6, 2003.

[HOYLE02] HOYLE D., RATTRAY M., JUPP R., BRASS A., Making Sense of MicroArray Data Distributions, *Bioinformatics*, **18**:576-584, 2002.

[IOSIFESCU97] IOSIFESCU D., SHENTON M., WARFIELD S., KIKINIS R., DENGLER J., JOLESZ F., MCCARLEY R., An automated registration algorithm for measuring MRI subcortical brain structures, *NeuroImage*, **6**: 13-25, 1997.

[JAIN02] JAIN A., TOKUYASU T., SNIJDERSA., SEGRAVES R., ALBERTSON D., PINKEL D., Fully Automatic Quantification of Microarray Image Data, *Genome Resarch*, **12**:325-332, 2002.

[JANG97] JANG J., LEE S., KIM S., Contour detection of hippocampus using dynamic contour model and region growing, *Proceedings of the 19th International Conference - IEEE/EMBS*, Chicago, 763-766, 1997.

[JOUENNE01] JOUENNE V., Critical Issues in the Processing of cDNA Microarray Images, *PhD thesis*, Virginia Polytechnic Institute and State University, 2001.

[JUNG02] JUNG H., CHO H., An automatic block and spot indexing with k-nearest neighbors graph for microarray image analysis, *Bioinformatics*, **18**:S141-S151, 2002.

[KAMBER95] KAMBER M., SHINGHAL R., COLLINS L., FRANCIS G., EVANS A., Model-based 3D segmentation of multiple sclerosis lesions in magnetic resonance brain images, *IEEE Transactions on Medical Imaging*, **14**:442-453, 1995.

[KAMITANI99] KAMITANI Y., SHIMOJO S., Manifestation of scotomas created by transcranial magnetic stimulation of human visual cortex, *Nature Neurosciences*, **2**:767-771, 1999.

[KAMITANI01] KAMITANI Y., BHALODIA V., KUBOTA Y., SHIMOJO S., A model of magnetic stimulation of neocortical neurons, *Neurocomputing*, **38**:697-703, 2001.

[KATAJAMMA03] KATAJAMMA M., Simulation Model for Exploring Variations in the Gene Expression Data and Its Analysis, *PhD thesis*, Université d'Helsinki, 2003.

[KAZTER03] KAZTER M., KUMMERT F., SAGERER G., A Markov Random Field Model of Microarray Gridding, *Proceedings of the 18th ACM Symposium on Applied Computing*, 2003.

[KELEMEN99] KELEMEN A., SZEKELY G., GERIG G., Elastic model-based segmentation of 3D neuroradiological data sets, *IEEE Transactions on Medical Imaging*, **18**:.828-839, 1999.

[KERR00] KERR K., MARTIN M., CHURCHILL G., Analysis of Variance for Gene Expression Microarray Data, *Journal of Computational Biology*, **7**: 819-837, 2000.

[KIKINIS96] KIKINIS R., SHENTON M., IOSIFESCU D., MCCARLEY R., SAIVIROONPORN P., HOKAMA H., ROBATINO A., METCALF D., WIBLE C., PORTAS C., DONNINO R., JOLESZ F., A digital brain atlas for surgical planning, model driven segmentation, and teaching, *IEEE Visualization and Computer Graphics*, **2**: 232-241, 1996.

[KIM01] KIM J., KIM H., LEE Y., A novel method using edge detection for signal extraction from cDNA microarray image analysis, *Experimental and molecular medicine*, **33**:83-88, 2001.

[KOHONEN97] KOHONEN T., Self Organized Maps, Springer Verlag, 1997.

[KOOPERBERG02] KOOPERBERG C., FAZZIO T., DELROW J., TSUKIYAMA T., Improved Background Correction for Spotted DNA Microarrays, *Journal of Computational Biology*, **9**:55-66, 2002.

[KOTALA01] KOTALA P., PERERA A., KAI ZHOU J., MUDIVARTHY S., PERRIZO W., DECKARD E., Gene Expression Profiling of DNA Microarray Data using Peano Count Trees, *Online Proceedings of the First Virtual Conference on Genomics and Bioinformatics*, October 2001.

[KOTHAPALLI02] KOTHAPALLI R., YODER S., MANE S. LOUGHRAN T., Microarray results: how accurate are they?, *BMC Bioinformatics*, **3**:22-31, 2002.

[KRINGS97] KRINGS T., BUCHBINDER B., BUTLER W., CHIAPPA K., JIANG H., COSGROVE G., ROSEN B., Functional magnetic resonance imaging and transcranial magnetic strimulation:

complementary approaches in the evaluation of cortical motor function, *Neurology*, **5**:1406-1414, 1997.

[KWAN96] KWAN R, EVANS A, PIKE G, An Extensible MRI Simulator for Post Processing Evaluation, *SPIE, Visualization in Biomedical Computing,* **1131**:135-140, 1996.

[LALUSH99] LALUSH D., Characterization, Modelling and Simulation of Mouse Microarray Data, *Critical Assessment of Microarray Data Analysis*, November 14-15, Durham, 2002.

[LAULUMAA93] LAULUMAA V, KUIKKA JT, SOININEN H, BERGSTROM K, LANSIMIES E, RIEKKINEN P., Imaging of D2 dopamine receptors of patients with Parkinson's disease using single photon emission computed tomography and iodobenzamide I 123, *Archives of Neurology*, **50**:509-512, 1993.

[LAWS03] LAWS R., BERGEMANN T., QUIAOIT F., ZHAO L., SignalViewer: analyzing microarray images, *Bioinformatics*, **19**: 1716-1717, 2003.

[LEE00] LEE M.T., KUO F.C., WHITMOREI G. A, SKLAR J., Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, *Proceedings of the National Academy of Sciences*, **97**: 9834-9839, 2000.

[LE MEUR01] LE MEUR N., Etude comparative de logiciels d'analyse d'image : extraction des données des puces à ADN, *Rapport de DEA Génomique et Informatique*, INSERM U533, Nantes, 2001.

[LEMKIN00] LEMKIN PF, THORNWALL GC, WALTON KD, HENNINGHAUSEN L., The microarray explorer tool for data mining of cDNA microarrays : application for the mammary gland, *Nucleic Acid Research*, **28**: 4452-4459, 2000.

[LI95] LI H., DE CUYPER D., HERMANUS A., NYSSEN E., CORNELIS J., Object recognition in Brain CT-Scans: knowledge-based fusion of data from multiple feature extractors, *IEEE Transactions on Medical Imaging*, **14**: 212-229, 1995.

[LIAO02] LIAO X., LAWRENCE C., Constrained Independent Component Analysis of DNA Microarrays Signals, *Proceedings of 1st Workshop on Genomic Signal Processing and Statistics*, Raleigh, USA, October 11-13, 2002.

[LIEW03] LIEW A., YAN H., YANG M., Robust adaptive spot segmentation ofDNA microarray images, *Pattern Recognition*, **36**:1251-1254, 2003.

[LOTUFO03] LOTUFO R., CHOUDHARY A., CORNELISON R., MOUSSES S., DOUGHERTY E., Nucleus Segmentation in Automated Cell Microarray Image Analysis, *Proceedings of the first International Conference on Bioinformatics and Computational Biology*,153, 2003.

[MACCABEE90] MACCABEE P., EBERLE L., AMASSIAN V., CRACCO R., RUDELL A., JAYACHANDRA M., Spatial distribution of the electric field induced in volume by round and figure-8 magnetic coils: relevance to activation of sensory nerve fibers, *Electroencephalography and Clinical Neurophysiology*, **76**:131-141, 1990.

[MACHL02] MACHL A.W., SCHAAB C., IVANOV I., Improving DNA array data quality by minimising neighbourhood effects, *Nucleic Acids Research*, **30**: e127, 2002.

[MCLACHLAN02] MCLACHLAN G., BEAN R., PEEL D., A mixture model based approach to the clustering of microarray expression data, *Bioinformatics*, **18**:413-422, 2002.

[MAGNOTTA99] MAGNOTTA V., HECKEL D., ANDREASEN A., CIZALDO T., CORSON P., EHRHARDT J., YUH W., Measurement of Brain Structures with artificial Neural Networks: two- and three-dimensional Applications, *Radiology*, **211**: 781-790, 1999.

[MARC02] MARC P, JACQ C., Arrayplot for visualization and normalization of cDNA microarray data, *Bioinformatics*, **18**:888-9, 2002.

[MARINO91] MARINO F., The passive propagation of electric field inside the head as modelled by the finite difference method, *Master thesis*, Université de Californie, 1991.

[MAZZIOTTA91] MAZZIOTTA C., VALENTINO D., GRAFTON S., BOOKSTEIN F., PELIZZARI C., CHEN G., TOGA A., Relating structure to function in vivo with tomographic imaging, *Ciba Foundation Symposium*, **163**: 93-112, 1991.

[MAZZIOTTA95] MAZZIOTTA J. TOGA A., EVANS A., FOX P., LANCASTER J., A probabilistic reference system for the human brain: theory and rationale for its development, *NeuroImage*, **2**: 89-101, 1995.

[MIRANDA03] MIRANDA P., HALLETT M., BASSER P., The electric field induces in the brain by magnetic stimulation: a 3D finite element analysis of the effects of tissue heterogeneity and anisotropy, *IEEE Transactions on Biomedical Engineering*, **50**:1074-1085, 2003.

[MONTAGNER02] MONTAGNER J., BUZER L., BARRA V., REVEILLES J.P., BOIRE J.Y., Utilisation de la géométrie discrète pour la fusion d'images anatomiques et fonctionnelles, *in Informatique et Santé,* SPRINGER, 173-180, 2002.

[MONTAGNER03] MONTAGNER J., BARRA V., BOIRE J.Y., Une approche géométrique pour la quantification de l'activité neuronale, *Actes du douzième forum des jeunes chercheurs en génie Biologique et Médical*, Nantes, 2003

[MOODY02] MOODY D., FADLIA B., SINGH A., SHAH S., MCINTYRE L., Quantitative comparison of image analysis software, in *DNA Array Image Analysis: Nuts & Bolts*, ed. Kamberova & Shah: 55-166, 2002.

[MORIO00] MORIO B, BARRA V, RITZ P, FELLMANN N, BONNY JM, BEAUFRÈRE B, BOIRE JY, VERMOREL M, Benefit of Endurance Training in elderly People over a short Period of Time is reversible, *European Journal of Applied Physiology*, **81**: 326-336, 2000.

[MOUSSAFIR92] MOUSSAFIR J.O., Voiles et polyèdres de Klein. Géométrie, algorithmes et statistiques, *Thèse, UFR Mathématiques de la Décision*, Paris IX, 1992.

[MUKHERJEE99] MUKHERJEE S, TAMAYO P, MESIROV JP,. SLONIM D, VERRI A, POGGIO T, Support vector machine classification of microarray data, *Technical Report* **182**: AI Memo 1676, CBCL, 1999.

[NADEEM03] NADEEM M., THORLIN T., GHANDHI O. PERSSON M., Computation of electric and magnetic stimulation in human head using the 3D impedance method, *IEEE Transactions on Biomedical Engineering*, **50**:244-255, 2003.

[NAGARAJAN03] NAGARAJAN R. Intensity-based segmentation of microarray images, *IEEE Transactions on Medical Imaging*, **22**:882-889, 2003.

[NETZU97] NETZU A., KIMURA S., OHTSUKI N., TANAKA M., Transcranial magnetic stimulation in benign childhood epilepsy with centro-temporal spikes, *Brain and Development*, **19**:134-137, 1997.

[NIEMANN95] NIEMANN K., POHL G., GRAF D., VON KEYSERLING G., Fuzzification of the Schaltenbrand and Wahren stereotaxic atlas, *proceedings of EUFIT'95*, 1648-1652, 1995.

[NOWINSKI01] NOWINSKI W., THIRUNAVUUKARASUU A., Atlas-assisted localization analysis of functional images, *Medical Image Analysis*, **5**:207-220, 2001.

[NUNEZ81] NUNEZ P., Electric fields of the brain, *Oxford University Press*, New York, 1981.

[NUNEZ90] NUNEZ P., Localization of brain activity with electroencephalography, *Advances in neurology,* **54**, 1990.

[O'ROURKE 98] O'ROURKE J., Computational geometry in C, 2nd edition, *Cambridge University Press*, 1998.

[PASCUAL94] PASCUAL-LEONÈ A., VALLS-SOLLE J., Resetting of essential tremor and postural tremor in Parkinson's disease with TMS, *Muscle and Nerve*, **17**:800-807, 1994.

[PAVLIDIS01] PAVLIDIS P, WESTON J, CAI J, GRUNDY WN, Gene functional classification from heterogeneous data, In RECOMB 2001: *Proceedings of the Fifth Annual International Conference on Computational Biology*, 249-255, 2001.

[PICKETT02] PICKETT S., BASARSKY T., VERDNIK D., WELLIS D., Microarray scanning and data acquisition; in Geschwind DH, Gregg JP (eds) *Microarrays for the Neurosciences: An essential guide*, MIT Press, Cambridge MA, 2002.

[PORTET01] PORTET F., TOUCHON J., CAMU W., Amyotrophic lateral sclerosis and cognitive disorders: review and analysis of the literature, *Revue Neurologique*, **157**:139-150, 2001.

[PREZANT87] PREZANT T., PFEIFER K., GUARENTE L., Organization of the regulatory region of the yeast cyc7 gene: multiple factors are involved in regulation, *Molecular Cell Biology*, **7**: 3252-3259, 1987.

[PRIDMORE99] PRIDMORE S., BELMAKER R., Transcranial magnetic stimulation in the treatment of psychiatric disorders, *Psychiatry and clinical Neurosciences*, **53**:541-548, 1999.

[PRUESSNER00] PRUESSNER J., LI L., SERLES W., PRUESSNER M., COLLINS L., KABANI N., LUPIEN S., EVANS A., Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories, *Cerebral cortex*, **10**: 433-442, 2000.

[PRUNIER03] PRUNIER C, PAYOUX P, GUILLOTEAU D, CHALON S, GIRAUDEAU B, MAJOREL C, TAFANI M, BEZARD E, ESQUERRE JP, BAULIEU JL., Quantification of dopamine transporter by 123I-PE2I SPECT and the noninvasive Logan graphical method in Parkinson's disease, *Journal of Nuclear Medicine,* **44**:663-670, 2003.

[RAMDAS01] RAMDAS L., COOMBES K.R., BAGGERLY K., ABRUZZO L., HIGHSMITH W.E., KROGMANN T., HAMILTON S.R. ZHANG W., Sources of nonlinearity in cDNA microarray expression measurements, *Genome Biology*, **2**:0047.1, 2001.

[RAMSAY97] RAMSAY J., SILVERMAN B., Functional Data Analysis, Springer-Verlag, 1997.

[RAPISARDA96] RAPISARDA G., BASTINGS E., DE NOORDHOUT A., PENNISI G., DELWAIDE P., Can motor recovery in stroke patients be predicted by early transcranial magnetic stimulation, *Stroke*, **27**:2191-2196, 1996.

[RAYCHAUDH00] RAYCHAUDHURI, S., STUART, J.M., ALTMAN, R.B., Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing*, **5**:452-463, 2000.

[REBAI01-A] REBAI H, LABORDE A, BARRA V, DELAÎTRE M, BOISGARD S, COUDEYRE L, COUDERT J, Effects of two electrical Stimulation Frequencies on thigh Muscles after knee Surgery, *Sixth Annual Congress of the European College of Sport Science*, Cologne, 2001.

[REBAI01-B] REBAI H, BARRA V, BONNY JM, POUMARAT G, COUDERT J, Thigh muscle volumes and proton T2 relaxation time after reconstruction of the anterior cruciate ligament and 12 weeks of rehabilitation, *Non Invasive investigation of muscle function Workshop*, Marseille, 2001.

[REBAI02] REBAI H, BARRA V, LABORDE A, BONNY JM, POUMARAT G, COUDERT J, Effects of two electrical stimulation Frequencies on thigh Muscles after knee Surgery, *International Journal of Sports Medicine*, **23** 604-9, 2002.

[REVEILLES01] REVEILLES J.P., The geometry of the intersection of voxels spaces, *Electronic Notes in Theoretical Computer Science, Elsevier Science*, **46**, 2001

[ROSENBLUM91] ROSENBLUM-VOS L.S., RHODES L., EVANGELISTA C.C. JR, BOAYKE K.A., ZITOMER R.S., The ROX3 gene encodes an essential nuclear protein involved in CYC7 gene expression in *Saccharomyces cerevisiae*, *Molecular Cell Biology*, **11**: 5639-47, 1991.

[ROSS00] ROSS D., SCHERF U., EISEN M., PEROU C., SPELLMAN P., IYER V., JEFFREY S., VAN DE RIJN M., WALTHAM M., PERGAMENSCHIKOV A., LEE J., LASHKARI D., SHALON D., MYERS T.,WEINSTEIN T., BOTSTEIN D., BROWN P., Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, **24**:227-235, 2000.

[ROTH90] ROTH B., BASSER P., A model for the stimulation of a nerve fiber by electromagnetic induction, *IEEE Transcactions on Biomedical Engineering*, **37**:588-597, 1990.

[ROWLAND01] ROWLAND L., SCHNEIDER A., Amyotrophic lateral sclerosis, *The New England Journal of Medicine*, **344**:1688-1700, 2001.

[RUDEMO02] RUDEMO M., LOBOVKINA T., MOSTAD P., SCHEIDL P., NILSSON S., LINDHAL P., Variance models for microarray images, *Report 2002:6*, Mathematical Statistics, Chalmers University of Technology, 2002.

[RUSH69] RUSH S., DRISCOLL D., EEG electrode sensitivity: an application of reciprocity, *IEEE Transactions on Bioledical Engineering*, **16**:15-22, 1969.

[SAEED02] SAEED N., PURI B., Cerebellum segmentation employing texture properties and knowledge based image processing: applied to normal adult controls and patients, *Magnetic Resonance Imaging*, **20**:425-429, 2002.

[SAPORTA91] SAPORTA G., Probabilités, analyse des données et statistique. *Ed. Technip*, 1991.

[SCHALTENB77] SCHALTENBRAND G., WAHREN W., Atlas for stereotaxy of the human brain, *2nd Ed. Thieme*, Stuttgart, 1977.

[SCHUCHHAR00] SCHUCHHARDT J, BEUKE D, MALIK A, WOLSKI E, EICKHOFF H, LEHRACH H, HERZEL H, Normalization strategies for cDNA microarrays, *Nucleic Acids Research* **28**:e47, 2000.

[SHENTON92] SHENTON M., KIKINIS R., JOLESZ F., POLLAK S., LEMAY M., WIBLE C., HOKAMA H., MARTIN J., METCALF D., COLEMAN M., MCCARLEY R., Abnormalities of the left temporal lobe and thought disorder in schizophrenia: a quantitative magnetic resonance imaging study; *New England Journal of Medicine*, **327**: 604-612, 1992.

[SIDDIQUI03] SIDDIQUI K., HERO A., SIDDIQUI M., Segmentation and Quantification of Microarray Images, *IEEE International Conference on Image Processing*, Barcelona, 2003.

[SONKA96] SONKA M., TADIKONDA S., Collins L., Knowledge-based interpretation of MR brain images, *IEEE Transactions on Medical Imaging*, **15**: 443-452, 1996.

[SPEED02] SPEED T.P., YANG Y.H., Direct versus indirect designs for cDNA microarray experiments, *The Indian Journal of Statistics*, **64**:706-720, 2002.

[SPELLMAN98] SPELLMAN P., SHERLOCK G., ZHANG M., IYER V., EISEN M., BROWN P., BOTSTEIN D., FUTCHER B., Comprehensive identification of Cell-Cycle-regulated genes of the Yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, **9**:3273-3297, 1998.

[STYNER03] STYNER M., GERIG G., LIEBERMAN J., JONES D., WEINBERGER D., Statistical shape analysis of neuroanatomical structures based on medial models, *Medical Image Analysis*, **7**:207-220, 2003.

[SZEKELY96] SZEKELY G., KELEMEN A., BRECHBUHLER C., GERIG G., Segmentation of 2-D and 3-D objects from MRI volume data using constrained elastic deformations of flexible Fourier contour and surface models, *Medical Image Analysis*, **1**:19-34, 1996.

[TALAIRACH88] TALAIRACH J., TOURNOUX P., Co-planar stereotaxic atlas of the human brain, *Georg Thieme Verlag*, Stuttgart, 1988.

[TAMAYO99] TAMAYO P., SLONIM D., MESIROV J., ZHU Q., KITAREEWAN S., DMITROVSKY E., LANDER E., GOLUB T., Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proceedings of the National. Academy of Sciences, 96:2907-2912, 1999

[TAVAZOIE99] TAVAZOIE S., HUGHES J., CAMPBELL M., CHO R., CHURCH G., Systematic determination of genetic network architecture, *Nature genetics*, **22**: 281-285, 1999.

[THOMPSON97] THOMPSON P., TOGA A., Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations, Medical Image Analysis, **1**:271-294, 1997.

[TIPPING99] TIPPING M., BISHOP C., Principal component analysis. *Journal of the Royal Statistical Society, Series*, **61**: 611–622, 1999.

[TOMIDA01] TOMIDA S., HANAI T., HONDA H., KOBAYASHI T., Gene expression analysis using fuzzy ART, *Genome Informatics*, **12**: 245-246, 2001.

[TÖRÖNEN99] TÖRÖNEN P, KOLEHMAINENB M, WONGA G, CASTRE E, Analysis of gene expression data using self-organizing maps, *FEBS Letters*, **451**:142-146, 1999

[TOURAILLE00] TOURAILLE E., BOIRE J.Y., Elastic registration of MRI scans using fast DCT, *World Congress on Medical Physics and Biomedical Engineering*, 2000.

[TSATSANIS03] TSATSANIS K., ROURKE B., KLIN A., VOLKMAR F., CICCHETTI D., SCHULTZ R., Reduced thalamic volume in high-functioning individuals with autism, *Biological Psychiatry*, **53**:121-129, 2003.

[TUZHILIN02] TUZHILIN A., ADOMAVICIUS G., Handling Very Large Numbers of Association Rules in the Analysis of Microarray Data, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 280-288, 2002.

[VAPNIK98] VAPNIK V., *Statistical Learning Theory*, Wiley, 1998.

[VILA96] VILA M, LEVY R, HERRERO M, FAUCHEUX B, OBESO J, AGID Y, HIRSCH E, Metabolic Activity of the Basal ganglia in parkinsonian Syndromes in Human and non-human Primates: a cytochrome oxidase histochemistry Study, *Neuroscience*, **71**:903-912, 1996

[VOGELS00] VOGELS O., VELTMAN J., OYEN W., HORSTINK M., Decreased striatal dopamine D2 receptor binding in amyotrophic lateral sclerosisand multiple system atrophy: D2 receptor down-regulation versus striatal cell degeneration, *Journal of Neurological Sciences*, **180**:62-65, 2000.

[WAGNER04] WAGNER T., GANGITANO M., ROMERO R., THEORET H., KOBAYASHI M., ANSCHEL D., IVES J., CUFFIN N., SCHOMER D., PASCUAL-LEONE A., Intracranial measurement of current densities induced by transcranial magnetic stimulation in the human brain, *Neurosciences Letters*, **354**:91-94, 2004.

[WANG02] WANG D., DODDRELL D., MR image-based measurement of rates of change in volumes of brain structures. Part I: method and validation, *Magnetic Resonance Imaging*, **20**:27-40, 2002.

[WANG01] WANG X., GHOSH S., GUO S., Quantitative quality control in microarray image processing and data acquisition, *Nucleic Acids Research*, **29**:e75, 2001.

[WARAGAI97] WARAGAI M., TAKAYA Y., HAYASHI M., Serial MRI and SPECT in amyotrophic lateral sclerosis: a case report, *Journal of Neurological Sciences*, **148**:117-120, 1997.

[WASSERMAN97] WASSERMANN E., GRAFMAN J., Combining transcranial magnetic stimulation and neuroimaging to map the brain, *Trends in cognitive sciencesI,* **6**:199-201, 1997.

[WATSON81] WATSON D., Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes, *The Computer Journal*, **24**:167-172, 1981.

[WIERLING02] WIERLING C., STEINFATH M., ELGE T., SCHULZE-KREMER S., AANSTAD P., CLARK M., LEHRACH H., RALF HERWIG R., Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis, *BMC Bioinformatics* **3**:29-45, 2002.

[WOODS92] WOODS R., CHERRY S., MAZZIOTTA J., Rapid automated algorithm for aligning and reslicing PET Images, *The Journal of Computer Assisted Tomography*, **16**: 620-633, 1992.

[WOOLF00] WOOLF P.J., WANG Y., A fuzzy logic approach to analyzing gene expression data, *Physiological Genomics*, **3**: 9-15, 2000

[WORTH98] WORTH A., MAKRIS N., PATTI M., GOODMAN J., HOGE E., CAVINESS J., KENNEDY D., Precise segmentation of the lateral ventricles and caudate nucleus in MR brain images using anatomically driven histograms, *IEEE Transactions on Medical Imaging*, **17**: 303-310, 1998.

[XING01] XING E., KARP R., CLIFF: Clustering of High-Dimensional Microarray Data via Iterative Feature Filtering Using Normalized Cuts, *Bioinformatics*, **1**:1-9, 2001.

[YANG00] YANG Y., BUCKLEY M., DUDOIT S., SPEED T., Comparison of methods for image analysis on cDNA microarray Data, *Technical Report*, **584**, University of Berkeley, 2000.

[YANG01] YANG Y., DUDOIT S, LUU P, SPEED T., Normalization for cDNA Microarray Data, *SPIE BiOS 2001*, San Jose, California, 2001

[YEUNG00-A] YEUNG KY, An empirical study on Principal Component Analysis for clustering gene expression data, *Technical Report* UW-CSE-2000-11-03, Department of Computer Science & Engineering, University of Washington, Seattle, 2000.

[YEUNG00-B] YEUNG, K.Y., RUZZO, W.L. Principal Component Analysis for clustering gene expression data, *Bioinformatics*, **17**:763-774, 2000.

[YEUNG01] YEUNG, K.Y., FRALEY C., MURUA A., RAFTERY A., RUZZO, W.L. Model-based clustering and data transformations for gene expression data, *Bioinformatics*, **17**:977-987, 2001.

[YOTSUTSUJI03] YOTSUTSUJI T., SAITOH O., SUZUKI M., HAGINO H., MORI K., TAKAHASHI T., KUROKAWA K., MATSUI M., SETO H., KURACHI M., Quantification of lateral ventricular subdivisions in schizophrenia by high-resolution three-dimensional magnetic resonance imaging, *Psychiatry Research: Neuroimaging,* **122**:1-12, 2003.

[YOUNG97] YOUNG M., TRIGGS W., BOWERS D., GREER M., FRIEDMAN W., Stereotactic pallidotomy lengthens the transcranial magnetic cortical stimulation silent period in Parkinson's disease, *Neurology*, **49**:1278-1283, 1997.

[YUE01] YUE H., EASTMAN P.S., WANG B.B., MINOR J., DOCTOLERO M.H., NUTTALL R.L., STACK R., BECKER J.W., MONTGOMERY J.R., VAINER M., JOHNSTON R, An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Research*, **29**:e29, 2001.

[ZADEH78] ZADEH L, Fuzzy Sets as a Basis for Theory of Possibility, *International Journal of Fuzzy Sets and Systems*, **1**:3-28, 1978

[ZUBAL96] ZUBAL I., High resolution, MRI-based, segmented, computerized head phantom, *Proceedings of the 6th International Radiopharmaceutical Dosimetry Symposium*, Gatlinburg, 319-324, 1996.

ANNEXES

Annexe A – Biologie des puces à ADN

Le concept de puce à ADN repose sur une technologie multidisciplinaire intégrant la biologie, la nanotechnologie, la chimie des acides nucléiques, l'analyse d'images et la bioinformatique. Grâce à cet outil, il est possible de mesurer le niveau d'expression de plusieurs milliers de gènes simultanément, et les applications (*e.g.* détermination des familles de gènes co-régulés, recherche de systèmes de régulation, …) sont en plein essor dans un grand nombre de domaines comme la pharmacologie, la médecine ou l'environnement.
L'analyse biologique des premières biopuces repose sur l'étude du transcriptome (étude des ARN messagers) d'une cellule ou d'un organisme. Si de nouvelles applications pour les puces à ADN ont également été développées (biopuces pour le diagnostic, pour la génomique comparative, pour l'identification de régions d'ADN régulatrices après ImmunoPrécipitation de la Chromatine (ChIPchips)), seules les biopuces transcriptome sont effectivement abordées dans ce chapitre.

1- De la cellule au transcriptome

De la cellule à l'ADN

La cellule est l'unité de base biologique. Chez les organismes les plus simples que sont les procaryotes (unicellulaires), le matériel génétique n'est pas compartimenté dans un noyau vrai mais est libre dans le cytoplasme. Pour les organismes plus complexes, les Eucaryotes (uni- ou pluricellulaires), l'information génétique est compartimentée dans un noyau (présence d'une membrane nucléaire). L'ADN, présent dans ce noyau, est une molécule qui peut être gigantesque avec un enchaînement linéaire de millions de nucléotides (l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T)). L'appariement des paires de nucléotides de type Watson et Crick (A et T d'une part, C et G d'autre part) concourt à l'organisation de l'ADN en double hélice, dont chaque brin a alors une séquence parfaitement complémentaire à l'autre (Figure A-1). Cette complémentarité est à la base de la plupart des techniques d'analyse en biologie moléculaire. Pour une espèce donnée, la séquence de la molécule d'ADN, *i.e.* l'enchaînement des nucléotides, est globalement semblable, sauf cas de polymorphismes.

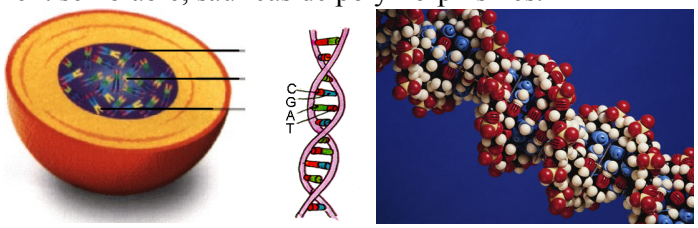


Figure A-1 : organisation de la molécule d'ADN

De l'ADN au génome

L'unité de base du stockage de l'information génétique est le gène, petite séquence d'ADN utilisée par la cellule pour synthétiser les protéines de structure, des enzymes nécessaires à son fonctionnement (respiration, division cellulaire, croissance, échanges, métabolismes…) ou des ARN (ARNt, ARNr et divers petits ARN). Les gènes sont de taille variable (de quelques centaines à plusieurs milliers de nucléotides), et sont disséminés le long du double brin d'ADN de chacune des cellules, souvent morcelés chez les organismes Eucaryotes en de nombreux fragments codants (exons), séparés par des séquences non codantes (introns). L'inventaire génique est alors défini comme l'ensemble de ces gènes, dont le nombre peut varier suivant l'individu, de 6200 chez la levure à une trentaine de milliers chez la souris ou l'homme. Le génome est donc de façon très simplifiée, l'ensemble de la séquence des chromosomes comprenant les gènes, des séquences répétées et des séquences fonctionnelles telles que les télomères (extrémités des chromosomes eucaryotes) ou les origines de réplications des chromosomes par exemple.

Du génome au transcriptome

En fonction de leurs besoins, les cellules utilisent à un instant donné une partie des gènes pour réaliser la synthèse des protéines nécessaires aux grandes fonctions cellulaires. Le passage du gène (contenu dans le noyau) à la protéine (synthétisée dans le cytoplasme) s'effectue en deux grandes étapes de transcription et de traduction, à l'aide d'un intermédiaire essentiel : l'ARN messager. La transcription correspond à la synthèse d'un brin d'ARN à partir de la double hélice d'ADN. L'ARN est synthétisé à partir du brin matrice (non codant) et a donc la même séquence que le brin non matrice (codant), la seule différence résidant dans le remplacement de la base T par la base U (uracile). Chez les Eucaryotes, l'ARN obtenu dans le noyau est appelé ARN pré-messager ou précurseur car il subira des maturations (élimination des introns : épissage, addition d'une queue polyA à l'extrémité 3' et d'une coiffe en 5') avant son transfert vers le cytoplasme (passage de la membrane nucléaire) où il sera traduit en protéines par les ribosomes. La traduction interprète chaque triplet de nucléotides (codon) de l'ARN messager en un acide aminé selon le code génétique universel : un codon donné correspond à un acide aminé donné, et plusieurs codons peuvent coder pour le même acide aminé (dégénérescence du code génétique). Ainsi, lors de la traduction, les ribosomes décodent l'information génétique pour permettre la synthèse de la protéine correspondante (Figure A-2). Au niveau des ribosomes, il existe une reconnaissance codon anti-codon entre la séquence de l'ARNm (codon) et la séquence anti-codon de l'ARNt (ARN de transfert) qui apporte l'acide aminé correspondant. Les ribosomes se déplacent sur l'ARNm pour permettre la synthèse complète de la protéine. La rencontre d'un codon non sens (aucun ARNt correspondant donc pas d'acide aminé) constitue le signal de fin de la traduction.
Signalons simplement que chez les Procaryotes, les gènes sont organisés en opérons (un ARNm porte plusieurs gènes), qu'il n'existe pas les mécanismes de maturation (absence d'introns) et enfin que la transcription et la traduction sont couplées du fin de l'absence de membrane nucléaire.

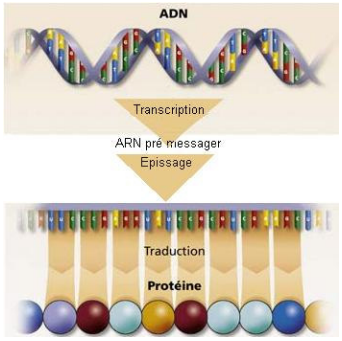


Figure A-2 : du gène à la protéine

Sachant qu'à partir d'un même gène, plusieurs copies d'ARN messagers peuvent être produits à des niveaux différents en fonction de l'activité de la cellule, le transcriptome reflétera donc le niveau d'expression de tous les gènes à un temps t pour une condition physiologique donnée. Il est le reflet instantané de l'activité cellulaire, et peut donc varier d'un type cellulaire à l'autre (neurone, cellule de la peau…), au cours du temps ou des différentes phases du cycle cellulaire, en fonction de conditions environnementales imposées aux cellules, ou en fonction de l'état sain ou pathologique de ces dernières. Si Jansen *et al.* [JANSEN01] ont montré que l'analyse du transcriptome donnait une assez bonne vision du jeu de protéines présent dans la cellule, il est cependant important de noter qu'il n'existe pas nécessairement de corrélation directe entre le niveau d'expression d'un gène (quantité d'ARNm) et la quantité de protéines produite. D'une manière plus générale, pouvoir comparer le transcriptome de différents types cellulaires, dans différentes conditions, ou pouvoir analyser l'ensemble du transcriptome d'une cellule à divers stades de son cycle cellulaire ou dans des conditions pathologiques, doit permettre d'une part de mieux comprendre le fonctionnement cellulaire sur le plan fondamental, et offre d'autre part beaucoup d'intérêts en termes d'applications potentielles.
L'analyse de l'abondance de chaque ARNm d'une cellule, d'un type tissulaire ou d'un organisme est depuis quelques années possible à grande échelle grâce à des techniques d'hybridation. Cette mesure

repose sur la propriété d'hybridation de l'ADN simple brin à son brin complémentaire de façon très stringente dans des conditions très contrôlées, et ce même si ce brin complémentaire n'est présent qu'en petites quantités au milieu d'autres fragments d'ADN. Depuis sa première apparition en 1975 (Southern Blot [SOUTHERN75]), la technique d'hybridation ADN-ADN généralement utilisée pour la cartographie des gènes a été adaptée à l'étude des ARN (Northern Blot). Dans la technique du Northern Blot, ce sont les ARN extraits qui après séparation sur un gel d'agarose dénaturant sont transférés sur une membrane (nitocellulose ou nylon). Par la suite, une sonde marquée spécifique du gène étudié permettra par hybridation complémentaire (ADN-ARN ou ARN-ARN) de caractériser le messager correspondant (taille, niveau d'expression). Ainsi, dans cette technique les cibles se trouvent sur le support solide et les sondes sont généralement marquées par de la radioactivité pour permettre la révélation de l'hybridation. Ces dernières années, cette technologie qui ne permettait l'étude de l'expression que d'un gène à la fois a évolué vers le développement d'un outil de post-génomique très puissant permettant de suivre de façon simultanée l'expression de tous les gènes d'une cellule ou d'un organisme donné. Il s'agit de la technologie très innovante des puces à ADN.

2- Les puces à ADN

Principe

Les puces à ADN [LANDER99] reposent sur le principe de complémentarité des brins de la double hélice d'ADN. Ce sont des supports de verre ou de silicium, de petite taille (de l'ordre de 25×75 mm) sur lesquels sont synthétisés directement ou greffés après synthèse (plus économique) des milliers de séquences d'ADN appelées sondes, caractéristiques d'autant de gènes. Les puces sont ensuite mises en contact avec l'ensemble des transcrits issus d'une cellule, les cibles, marqués par une molécule fluorescente (fluorochrome). Le mécanisme d'hybridation implique que les cibles marquées reconnaissent, parmi les sondes de la puce, celles qui leur sont complémentaires, et s'y apparient. Après cette étape, les hybrides sondes/cibles sont imagés *via* un scanner, repérés et quantifiés grâce à leur fluorescence. En déposant un nombre important de sondes, on peut ainsi analyser un grand nombre de gènes simultanément. On peut même, en utilisant deux fluorochromes tels que les cyanines (*e.g.* un rouge Cy5 et un vert Cy3) comparer les niveaux d'expression relatifs de deux transcriptomes différents sur une seule puce (sain/pathologique, début/fin de cycle cellulaire,…) : pour chaque fluorochrome, deux images distinctes sont obtenues et superposées *in silico* pour analyser le différentiel d'expression (Figure A-3).

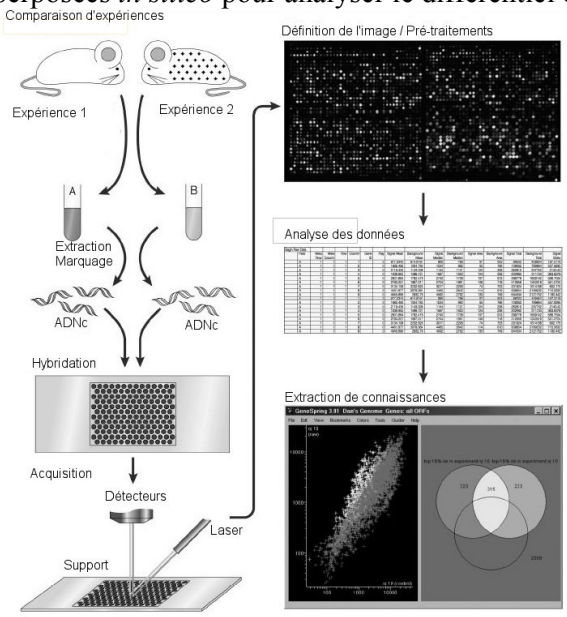


Figure A-3 : schéma de principe des expériences à base de puces à ADN

La préparation des sondes

Selon l'application désirée et le modèle biologique étudié, plusieurs types de biopuces peuvent être utilisées :

- étude du transcriptome : biopuces ADNc ou oligonucléotides :
- diagnostic : biopuces oligonucléotides ;
- génomique comparative, ChIPchips : biopuces ADN génomique.

Dans le cadre d'une étude du transcriptome, les sondes peuvent être des séquences d'ADN double brin d'une taille en moyenne de 400 paires de bases (ADNc, produits PCR) ou simple brin (oligonucléotides généralement de 50, 60 ou 70 bases, avec des puces Affymetrix utilisant de petits oligonucléotides de 20 bases).

La conception des biopuces de type ADNc nécessite une première étape d'obtention des sondes grâce à une amplification PCR (Polymerase Chain reaction) à partir d'une collection d'ADNc ou directement sur l'ADN génomique s'il s'agit d'Eucaryotes simples sans introns ou de Procaryotes. Un contrôle de ces produits d'amplification doit être effectué (taille, séquence) pour être certain qu'il s'agit de la bonne sonde. L'ADN des sondes est dénaturé pour le dépôt en simple brin pour permettre par la suite une hybridation avec les cibles marquées. Les sondes ainsi préparées sont déposées par un robot (le spotter) sur le substrat de la puce, recouvert d'une fine couche de polymère (généralement du γ-amino-propyl-silane) permettant la fixation des sondes par simples liaisons électrostatiques. Le dépôt, ou spot, est réalisé à l'aide d'aiguilles (creuses ou par la technique de l'anneau pin and ring) sous la forme d'une gouttelette de 2 nl environ. Le diamètre des spots peut varier de 80 à 300 µm, et la distance entre deux dépôts consécutifs sur la lame est de l'ordre de 250 µm dans les deux directions.

Concernant les biopuces oligonucléotides, la synthèse des sondes est chimique ce qui évite les premières étapes laborieuses de PCR et de contrôles rencontrées lors de la conception des biopuces ADNc. Ces oligonucléotides joueront donc le rôle « d'étiquettes » spécifiques identifiant les différents gènes.

Les puces Affymetrix (http://www.affymetrix.com) très haute densité utilisent une technologie permettant une synthèse directe, *in situ,* des oligonucléotides sur le support solide (semi-conducteurs, masquage et photolithographie). Ainsi, 65536 oligomères peuvent être obtenus sur 1,6 cm^2 en 32 étapes.

Pour les autres biopuces oligonucléotides, la synthèse se fait avant le dépôt sur le support solide. Plusieurs modes de dépôt peuvent être rencontrés, soit par contact direct (aiguille creuse ou pin and ring) soit par la technique « jet d'encre » (biopuces Agilent).

La préparation des cibles et l'hybridation

Des milliers de transcrits différents sont présents dans les cellules à un moment donné, et leur abondance relative est révélatrice de l'activité cellulaire à cet instant. La préparation des cibles consiste tout d'abord à extraire du milieu cellulaire ces transcrits puis leur intégrer un marqueur fluorescent (rouge ou vert) qui permettra d'évaluer et de quantifier l'appariement sonde/cible. Le marquage, soit direct (incorporation d'un nucléotide fluorescent), soit indirect (incorporation d'un nucléotide modifié amino allyl dUTP puis couplage du flurochrome dans un second temps), se fait suite à la transcription inverse des ARNm permettant l'obtention d'un brin d'ADNc fluorescent. Du fait de la complémentarité des nucléotides, le dépôt des cibles marquées sur la puce déclenche l'appariement des séquences sondes/cibles complémentaires. Cette hybridation, qui dure quelques heures en milieu liquide, est suivie d'un lavage du substrat qui permet d'éliminer les cibles non fixées, ou fixées non spécifiquement. Après séchage la puce est passée au scanner pour repérer les hybridations.

La chaîne d'acquisition de l'image de la puce

La puce est alors révélée par un lecteur (scanner) muni de lasers qui permettent d'exciter les molécules de fluorochromes et de détecter par un microscope confocal le signal émis dans chaque spot. Dans le cas du marquage avec deux fluorochromes (vert Cy3 et rouge Cy5), une image numérique est acquise pour l'échantillon marqué avec le Cy3 et une en Cy5. Un spot de couleur verte indique un gène dont le niveau d'expression est plus élevé dans l'échantillon marqué avec le Cy3 que celui marqué avec le Cy5, et inversement pour un spot de couleur rouge. Le spot apparaît jaune lorsque le gène est exprimé de manière

identique dans les deux échantillons comparés. L'analyse des données numériques issues de l'acquisition est effectuée par un logiciel qui prétraite, segmente et quantifie les différents niveaux d'activité dans les spots.

3- L'analyse des images de puces

L'analyse d'image est un aspect central des expériences menées à l'aide des puces à ADN. Dans le cas d'un double marquage, le but de cette phase est de quantifier, de manière relative, le niveau d'expression des gènes. Cette mesure, basée sur un rapport d'intensité entre les deux niveaux des fluorochromes détectés, est fonction de nombreux paramètres dépendant des méthodes utilisées, des mesures expérimentales et des conditions biologiques. Globalement, cette étape d'analyse a un impact considérable sur l'interprétation biologique des données, et repose sur quatre phases (Figure A-4).

La localisation des spots

Il s'agit ici de déterminer les coordonnées de chaque spot de la puce. Cette étape est normalement effectuée à l'aide d'une grille théorique définie lors du plan de dépôt des sondes. Pour localiser un spot sur une image, c'est-à-dire faire correspondre un modèle idéal de puce avec une image acquise, un nombre important de paramètres doit être estimé (espaces entre les spots, espaces entre blocs d'une puce…). De même, les mouvements de translation des spots ou des blocs de spots, liés à la variation de position des aiguilles du spotter, sont à évaluer. Le repérage des spots doit être simple et rapide à réaliser. L'automatisation de cette tâche permet une accélération considérable de l'analyse. Enfin, la définition des grilles doit être aussi juste que possible. En effet, de la précision de cette étape dépend l'efficacité des mesures ultérieures.

La segmentation des spots

Il s'agit ici de classer les pixels de l'image en deux classes, "fond" et "signal". Ceci sous-tend une analyse du signal au niveau de chaque spot et un découpage de l'image en différentes régions, chacune ayant des propriétés propres. En raison des propriétés biologiques des dépôts et de la physique de la chaîne d'acquisition, le bruit acquis est fortement non uniforme dans l'image et varie en fonction d'un certain nombre de paramètres. De plus, à ce stade de l'analyse, la taille variable, la forme complexe et les irrégularités des spots compliquent la tâche de segmentation, mais doivent cependant être déterminés de manière semi-automatique voire automatique.

L'extraction d'indices

Maintenant que les pixels "signal" sont identifiés, il s'agit d'extraire un ou plusieurs indices pertinents permettant de quantifier le niveau d'expression de chaque gène. Ce niveau correspond à une mesure relative des intensités de fluorescence en rouge et vert. Un filtrage des données permet de sélectionner des variations significatives (tri des spots sur un barème de critères), à partir desquelles un rapport de fluorescences est calculé. La valeur de ce ratio indique l'induction ou la répression du gène.
Dans le cas de la comparaison de données issues de plusieurs expériences (données temporelles, sain/pathologique…), il faut de plus nécessairement normaliser les données avant quantification pour éliminer les artefacts dus par exemple au protocole expérimental.

La classification des données

Enfin, les données préparées par les étapes précédentes permettent de regrouper les gènes par familles ayant des comportements semblables, en analysant leurs niveaux d'expression. Cette étape de classification permet de répondre à des problématiques biologiques importantes allant de la recherche de gènes corégulés au cours du développement ou en réponse à des perturbations environnementales, à l'identification des liens entre des variations d'expression et d'autres données biologiques en passant par la recherche automatique de systèmes de régulation.

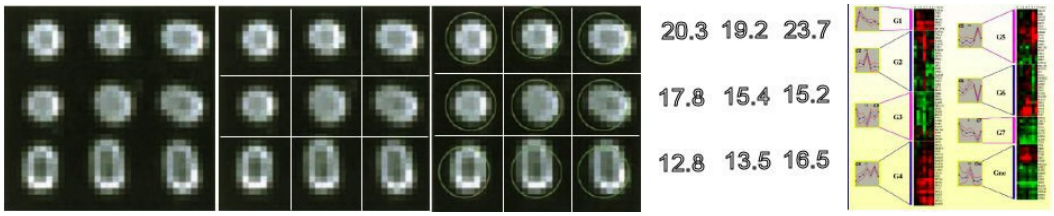


Figure A-4 : schéma récapitulatif des étapes de traitement des images de puces à ADN

4- Les applications

L'utilisation des outils d'analyse des puces à ADN est aujourd'hui de plus en plus intensive (environ 12500 articles parus en 2002 sur l'étude du transcriptome [Scheel02]), et ceci dans de nombreux domaines (Figure A-5), parmi lesquels :

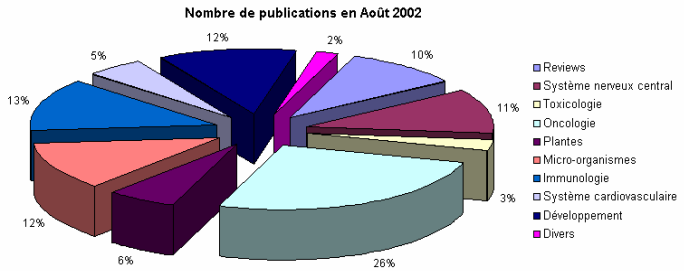


Figure A-5 : domaines de publication des études sur les puces à ADN.

Etude de la dynamique du transcriptome

Bien sûr, un des premiers rôles des puces à ADN a été d'étudier l'expression de milliers de gènes en parallèle, souvent selon plusieurs conditions. De très nombreuses études ont été réalisées sur divers organismes (*e.g.* la levure *Saccharomyces cerevisiae* [SPELLMAN98], *Streptomyces coelicolor* [HUANG01] dont dérivent 60% des antibiotiques actuellement utilisés), en fonction de diverses conditions expérimentales (cycle cellulaire, influence du milieu extérieur,…), pour mettre en évidence des gènes co-régulés, des liens entre des variations d'expression et d'autres données biologiques ou des systèmes de régulation.

Nous collaborons actuellement avec une équipe de biologistes de l'Université Blaise Pascal de Clermont-Ferrand (laboratoire de biologie des protistes, équipe Génomique Intégrée des Interactions Microbiennes, Pierre Peyret, UMR 6023 CNRS) sur une telle étude, concernant le pathogène *Encephalitozoon cuniculi* [KATINKA01].

Recherche et action de médicaments

Les puces ADN sont un outil de choix dans la recherche et la caractérisation de nouvelles molécules à visée thérapeutique [DEBOUK99]. En effet, la plupart des médicaments agissent en inhibant leur molécule cible, et une mutation du gène correspondant devrait donc avoir un effet similaire sur le transcriptome de la cellule. La société Affymetrix propose ainsi depuis fin 1997 une puce de 800 oligonucléotides permettant la détection des 18 mutations et des 10 polymorphismes répertoriés dans deux gènes associés à un système enzymatique (cytochrome P450) qui joue un rôle essentiel dans la cinétique de dégradation par le foie de nombreux médicaments. De même, Marton *et al.* [MARTON98] ont utilisé une puce à ADN contenant l'ensemble des gènes de la levure afin de démontrer l'existence d'une corrélation significative entre le profil obtenu lors d'une stimulation médicamenteuse antimicrobienne et le profil d'expression d'une levure portant un gène muté et impliqué dans le métabolisme d'action de ce médicament. Ce

principe peut être exploité pour la création d'une base de données contenant un grand nombre de profils d'expression, provenant à la fois de cellules stimulées par des médicaments et de souches contenant différentes mutations. Ces données offrent un moyen de décoder les profils complexes d'expression de groupes de gènes modulés par différentes classes de médicaments.
Dans ce domaine, les perspectives à moyen terme vont du diagnostic prédictif à la thérapie personnalisée, comme précisé dans un rapport récent [PROVENCE02] de la Direction générale de l'Industrie, des Technologies de l'information et des Postes sur la société de l'information (Figure A-6)

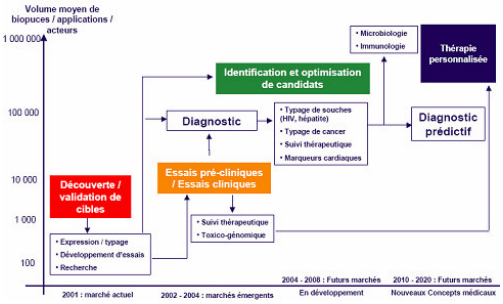


Figure A-6 : évolution des puces à ADN dans les applications pharmaceutiques.

Recherche de mutations et de remaniements chromosomiques

Pour chaque maladie héréditaire, la palette des mutations possibles est en général très étendue, ce qui rend la recherche de la mutation responsable lors du diagnostic très complexe. De nombreuses techniques de criblage ou de caractérisation de mutations connues ont été développées, mais restent aujourd'hui peu performantes. Les techniques de criblage permettent de rechercher une variation de séquence, caractérisée ensuite par séquençage. Une recherche directe de la mutation est aussi possible par séquençage, mais son coût est encore très élevé. La recherche des mutations est donc par excellence un champ d'activité où les puces apportent une révolution [CRONIN96]. De même, certains auteurs proposent de dédier des puces à ADN à l'étude de remaniements chromosomiques (*e.g.* aneuploïdie), par exemple pour classer différents types de rhabdomyosarcomes (plus fréquente des tumeurs des tissus mous en pédiatrie, qui représente 10% des tumeurs solides chez l'enfant) [LU01].

Oncologie

L'exemple précédent illustre un des principaux intérêts des puces à ADN pour la classification des tumeurs, le diagnostic et le pronostic en cancérologie. Des travaux ont plus généralement mis à jour une signature d'expression fortement associée au risque d'évolution métastatique dans le cancer du sein [Van't VEER02], en étudiant le profil d'expression de tumeurs primaires de ce cancer. Cette même pathologie a de même été étudiée par Mei *et al.* [MEI00], qui proposent de détecter à l'aide des puces à ADN haute densité des allèles dangereux du gène de susceptibilité au cancer du sein BRCA1.

Environnement et agriculture

Dans une chaîne de production agroalimentaire, les puces à ADN sont des outils particulièrement efficaces pour reconnaître l'origine des espèces animales qui composent les produits alimentaires, mais également pour détecter et contrôler dans des semences des séquences provenant d'organismes génétiquement modifiés [KUIPER01]. Les applications dans l'environnement concernent quant à eux l'analyse microbiologique de l'eau de consommation, et plus généralement le contrôle qualité en milieu industriel, pour renforcer la sécurité des procédés et des produits. Ainsi, la Lyonnaise des Eaux et bioMérieux ont par exemple décidé d'unir leurs compétences pour mettre au point, dans le cadre du projet Aquagen, une technique d'analyse de l'eau potable par puces à ADN, qui apporte aux consommateurs une garantie renforcée en matière de contrôle de la qualité de l'eau.

Bibliographie

[CRONIN96] CRONIN M, MIYADA C, FUCINI R, KIM S, MASINO R, WESPI R., Detecting cystic-fibrosis mutations by hybridization to DNA-probe, *Biologicals*, **24**:209, 1996.

[DEBOUK99] DEBOUK C., GOODFELLOW P., DNA microarrays in drug discovery and development, *Nature Genetics*, **1**: 48–50, 1999.

[HUANG01] HUANG J., LIH C., PAN K., COHEN S., Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in Streptomyces coelicolor using DNA microarrays, *Genes Development*, **15**:3183-3192, 2001.

[JANSEN01] JANSEN R., GREENBAUM D., GERSTEIN M., Relating whole-genome expression data with protein-protein interactions, *Genome Research*, **12**:37-46, 2002.

[KATINKA01] KATINKA M., DUPRAT S., CORNILLOT E., MÉTÉNIER G., THOMARAT F., PRENSIER G., BARBE V., PEYRETAILLADE E., BROTTIER P., WINCKER P., DELBAC F., EL ALAOUI H., PEYRET P., SAURIN W., GOUY M., WEISSENBACH J., VIVARÈS C., Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature*, **414**:450-453, 2001.

[KUIPER01] KUIPER H., KLETER G., NOTEBORN H., KOK E., Assessment of the food safety issues related to genetically modified foods, *Plant Journal*, **27**:503-528, 2001.

[LANDER99] LANDER E., Array of hope, *Nature Genetics*, **21**:3-4, 1999.

[LU01] LU Y., WILLIAMSON D., CLARK J., WANG R., TIFFIN N., SKELTON L., GORDON T., WILLIAMS R., ALLAN B., JACKMAN A., COOPER C., PRITCHARD-JONES K., SHIPLEY J., Comparative expressed sequence hybridization to chromosomes for tumor classification and identification of genomic regions of differential gene expression, *Proceedings of National Academy of Sciences*, **98**:9197-9202, 2001.

[MARTON98] MARTON M., DERISI J., BENNETT H., IYER V., MEYER M., ROBERTS C., STOUGHTON R, BURCHARD J, SLADE D, DAI H, BASSETT D. JR, HARTWELL L., BROWN P., FRIEND S., Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine*, **4**:1293-301, 1998.

[MEI00] MEI R., GALIPEAU P., PRASS C., BERNO A., GHANDOUR G., PATIL N., WOLFF R., CHEE M., REID B., LOCKHART D., Genome-wide detection of allelic imbalance using human SNPs and high density DNA arrays, *Genome Research*, **10**:1126-1137, 2000.

[PROVENCE02] PROVENCE M., Etude biocapteurs et biopuces, *Rapport de la Direction générale de l'Industrie, des Technologies de l'information et des Postes sur la société de l'information*, 2002.

[SCHEEL02] SCHEEL J, VON BREVERN MC, HORLEIN A, FISCHER A, SCHNEIDER A, BACH, A. Yellow pages to the transcriptome, *Pharmacogenomics*, **3**:791-807, 2002.

[SOUTHERN75] SOUTHERN E., Detection of specific sequences among DNA fragments separated by gel electrophoresis, *Journal of Molecular Biology*, **98**:503-517, 1975.

[VAN'T VEER02] VAN'T VEER, L., Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, **415**: 530-536, 2002.

Annexe B - Dépôt de brevet

Le texte proposé sert de base à la demande de brevet déposée par la société SEGAMI SA pour la valorisation du processus de quantification complet développé dans le cadre de la thèse d'Emmanuelle Frenoux, effectuée en contrat CIFRE avec cette société.

A new method for the quantitative study of neurotransmission

Inventors : Emmanuelle FRENOUX, Vincent BARRA, Arnaud COLIN
(during PhD studies under the supervision of Jean-Yves BOIRE)
ERIM (CENTI), Faculty of Medicine, BP 38, 63001 Clermont-Ferrand Cedex, France.

We propose a new method for the automatic quantification of neurotransmission using data fusion. The process uses a Magnetic Resonance image as anatomical reference and a Single Photon Emission Computed Tomography image of the same patient for functional information. Once both acquisitions are coregistered, the process is divided in two steps: first, structures of interest are automatically segmented using a data fusion-based algorithm, then the masks obtained are used to compute functional indexes in the SPECT. Both steps of the process are now validated and we plan to apply it as a tool for differential diagnosis in a wide range of pathologies.

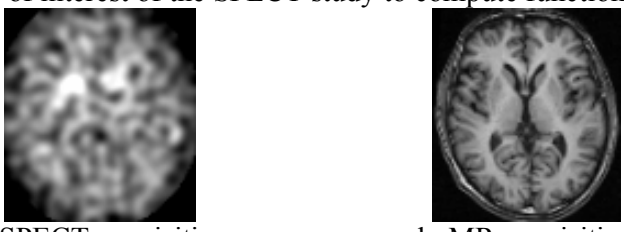
I. INTRODUCTION

Several pathologies are induced by neurotransmission process dysfunctions leading to severe troubles and possible death (e.g. Parkinson's disease). In order to study these pathologies, functional studies are acquired using Single Photon Emission Computed Tomography (SPECT) imaging and specific radioactive tracers allowing the quantification of neurotransmission efficiency [ELFA-01, HABR-99].

This kind of study has a poor spatial resolution (7-10 mm, see Fig. 1.a) which hinders thin cerebral structures location. To obtain a more reliable quantitative study, many clinicians simultaneously acquire a Magnetic Resonance (MR) 3D image of the same patient. This kind of acquisition provides good anatomical information with a very high spatial resolution (voxels are about 1 mm wide, see Fig. 1.b).

Both acquisitions are used by the clinician to quantify neurotransmission process, cerebral structures of interest are manually outlined on the MR image and the result is superimposed to the SPECT study for functional indexes computation [CATA-01].

We propose a new method to automatically locate and quantify neurotransmission using a SPECT study and a MR image of the same patient. Once both acquisitions are registered, the process is divided in two steps: first the structures of interest are automatically segmented on the MR image; the resulting binary mask is then used as anatomical reference to compute volumetric indexes from the MR image and to precisely locate the regions of interest of the SPECT study to compute functional indexes.



a. SPECT acquisition b. MR acquisition

Fig. 1. SPECT and MR acquisitions.

II. METHODOLOGY

First, MR and SPECT images have to be put in the same geometric referential, i.e. registered. This part can be achieved with tools previously developed by SEGAMI Corporation and available on Mirage system.

A. Segmentation of cerebral structures

The main cerebral structures implied in dopaminergic neurotransmission are the heads of caudate nuclei and the putamens (Fig. 2).

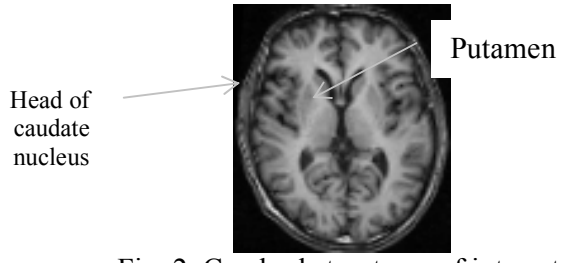


Fig. 2. Cerebral structures of interest.

The cerebral structures of interest were extracted using a process fusing numerical information extracted from the MR image and contextual information (direction, distance, shape, …) provided by a clinician. Each piece of information is represented by a fuzzy set, using possibility theory and fused to the others. The whole process is summarized in Fig. 3 and has already been assessed and published [FREN-01].

1. Introduction

Cerebral structures segmentation in medical imaging has numerous clinical applications. It can provide assistance tools for pathologies forecast [BOU-96] and follow up [SCHU-99]. It can also be used as an help to surgery and radiotherapy [BARR-00] or to obtain an anatomical reference for functional studies [COLI-99].

Various segmentation methods are inventoried in literature, many of them requiring an operator intervention. For example, region growing [ZHU-95] for tumors detection or deformable contours [ASHT-95, JANG-97] for hippocampus segmentation need to be initialized. In [DHAW-96, HEIN-92] interactive methods using mathematical morphology are proposed; other methods (e.g. neural networks [OZKA-94] or a modified k-nearest neighbors rule [VINI-95]) require a learning step. Finally, some segmentation methods are fully automatic. For example those using data fusion to aggregate information stemming from images (numerical data) [BARI-94], or theoretical knowledge and numerical data [BARR-00, HILT-01]. Géraud [GERA-98], in particular, proposes a segmentation method using anatomical knowledge and information extracted from an atlas.

We propose here to mimic the way the clinician looks for a cerebral structure in an MRI using an automatic segmentation method. He synthesizes the information brought by the image and his own knowledge (shape, matter, distance, direction) to locate the structure. The segmentation scheme is divided in three steps: first the representation of numerical (image) and contextual (expert) information in the same theoretical frame, then its fusion and last the decision step.

2. Cerebral structures of interest

The method is illustrated with the segmentation of putamens (P) and heads of caudate nuclei (HCN). These structures are affected by numerous diseases such like Parkinson's disease or schizophrenia.

Caudate nuclei (CN) are gray matter coma-shaped structures coiling up the thalami and going down behind them. The HCN is ovoid, rather bulky and bulges into the lateral wall of the lateral ventricles (LV) frontal horn. P are pyramidal-shaped gray matter structures and constitute the side part of the lenticular nuclei. P and CN carry out, among others, motor functions. Fig. 1 shows these structures of interest.

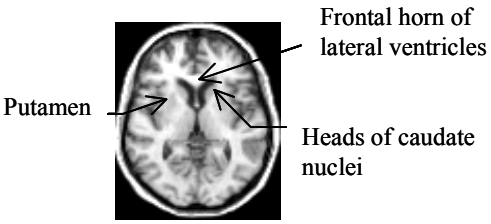


Fig. 3. View of the structures of interest on a T1-weighted MRI

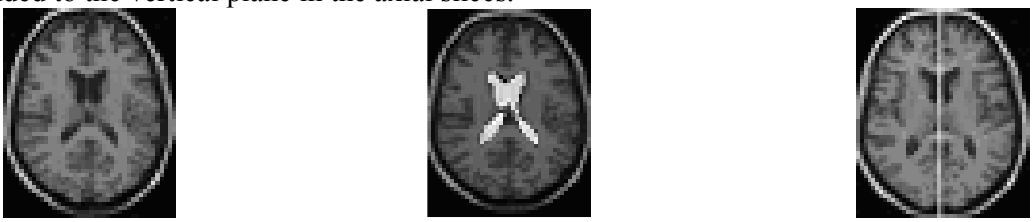
Theoretical knowledge concerning these structures has been collected from an expert and represented within the same theoretical framework. It has then been fused in order to segment P and HCN in MR images.

3. Modeling and fusion of information

Data fusion is defined here as an aggregation of conflicting, ambiguous, supplementary and/or redundant information, allowing more accurate or less uncertain data interpretation. Fusion has to manage uncertainties and inaccuracies, like a specialist does while observing several medical images, to avoid inconsistencies.
While modeling information, possibility theory [ZADE-78, DUBO-88] allows taking into account the fact that shape and volume of the structures vary from one subject to another according to his age, sex and pathologies. It is possible to segment structures of interest using reference structures which can easily be spotted (called landmarks). Information was provided by an expert in addition to data which is extracted from the MRI. Each piece of information is modeled as a fuzzy map to be fused.
Four kinds of data are modeled and fused to extract cerebral structures : numerical information extracted from de treated MRI, direction, distance and shape, modeled from the expert's anatomical knowledge in the repair of the treated image. Data representation is divided in five steps : 1-extraction of numerical data, 2-Landmarks extraction, 3-Direction representation with respect to the landmarks, 4-Distance representation with respect to the landmarks, 5-Shape information introduction.

Numerical information extracted from MRI: five or four tissue classes (background, cerebrospinal fluid (CSF), white matter, gray matter and if necessary subcutaneous fat) are extracted from the MR image using fuzzy C-means possibilistic clustering algorithm. This algorithm creates five or four fuzzy “matter maps”, depending on image contrast, in which one voxel gray level represents its membership to the considered tissue.

Segmentation of the landmarks: Fuzzy maps were then used to segment the anatomical landmarks. The frontal horn of LV and the inter-hemispheric plane (Fig. 2) are the landmarks used to model contextual information. LV were extracted from a binary CSF map (obtained by thresholding the fuzzy CSF map) using mathematical morphology operations. The rough location of the inter-hemispheric plane was then calculated by maximizing Pearson’s correlation coefficient between the two halves of the image. The patients were supposed to be always placed in the MR scan so that the inter-hemispheric plane roughly corresponded to the vertical plane in the axial slices.

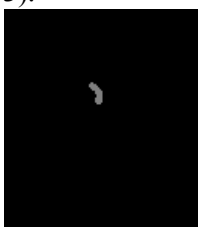
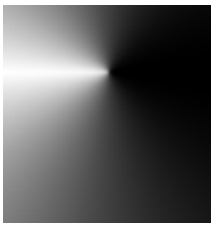


MRI LV Inter-hemispheric plane

Fig. 4. Results of landmarks segmentation on an axial slice

Information concerning directions: we model by a fuzzy set a vague sentence like “the structure S1 is in the direction D with respect to S2” where S2 is a structure already segmented. D is represented in

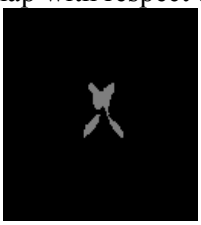
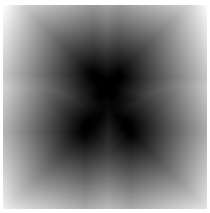
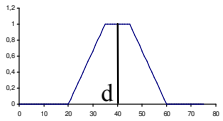
spherical coordinates and we use fuzzy mathematical morphology [BLOC-95] to obtain a fuzzy map in which one voxel gray level represents its membership to the domain "in direction D with respect to S2" [GERA-98] (Fig. 3).

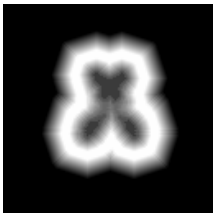



Frontal horn of the left LV "Rather on the left of the frontal horn of the left LV"

Fig. 5. Example of fuzzy direction map

Information concerning distances: The piece of information modeled here is a vague sentence like "the structure S1 is at distance F(d) from S2" where F(d) is a linguistic modifier ("almost", "inferior to", "superior to") applied to distance d. We use the method described in [BARR-00] to create the fuzzy distance map with respect to S2 (Fig. 4).



LV Chamfer map "Almost at d" "Almost at d from LV"

Fig. 6. Example for distance modelling

Representation of shape information

To create a fuzzy model of structures shapes, we used a binary segmentation of P and HCN on 48 co-registered MR images. In this model (Fig. 5), one voxel gray level represents its frequency of appearance in the considered structure. During the segmentation process, this map is registered on the MRI and warped on it using the algorithm proposed by Touraille et al [TOUR-00].

Fig. 7. Example of fuzzy shape map

Data fusion and decision step: for the fusion step, we use two operators: max (for complementary information) and min (for redundant information) operators. The fusion step results in a fuzzy map in which gray levels are the memberships to the required structure with respect to the whole set of numerical and contextual data. The last step is the decision step. Only surest voxels are conserved in interaction with the user.

The whole fusion process is summarized in Fig. 8.

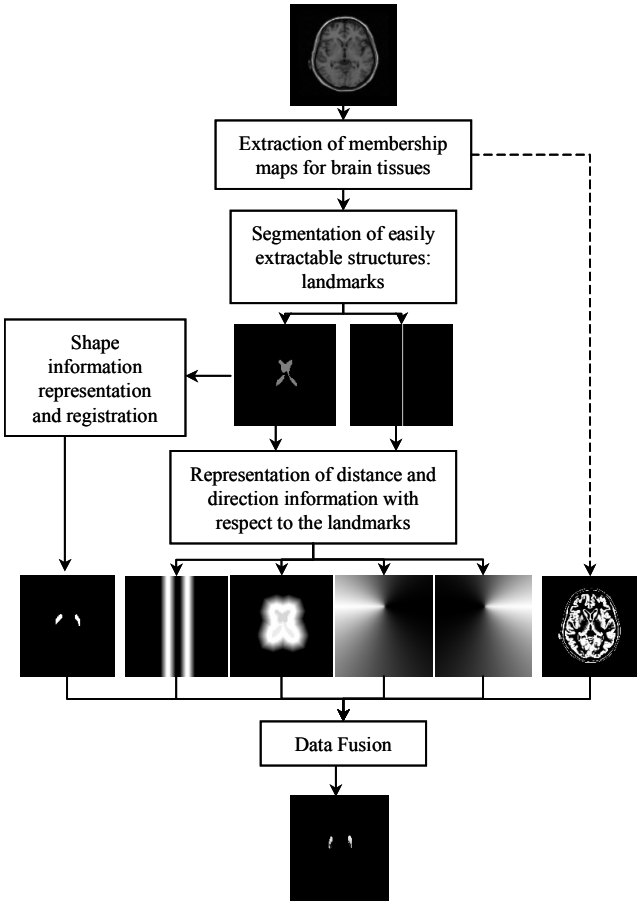
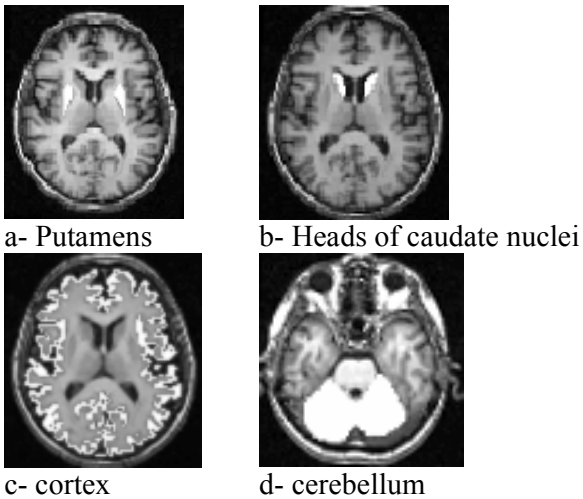


Fig. 8. Data fusion process for putamens segmentation

The result of segmentation process for heads of caudate nuclei and putamens is presented on Fig. 9., superimposed with the corresponding MRI.



a- Putamens b- Heads of caudate nuclei

c- cortex d- cerebellum

Fig. 9. Result of the segmentation process

This segmentation process has already been assessed and published [FREN-01]. It can be applied to extract any cerebral structure provided that it can be described by spatial, shape and matter information and that contrast and resolution of the MRI are sufficient. It is already applied on Mirage system to

extract heads of caudate nuclei, putamens, cerebellum, lateral ventricles, and parts of the cortex. The method for thalamus segmentation is being developed.

B. Quantification step
Statistical indexes usually computed in literature for quantification can be divided in two classes: volumetric indexes, computed from the MR image alone, and binding indexes, computed from both MR and SPECT acquisitions.

1. Volumetric indexes

Several volumetric indexes are computed from the binary mask resulting of the segmentation process: the volume of the structures, the volume of the structures normalized by the whole brain volume, a volumetric asymmetry index I=L/R, where L (resp. R) is the volume of left (resp. right) structure; and an absolute asymmetry coefficient, A, computed as:
A=100%*|L-R|/(0.5*(L+R)) (where L and R were the same as described above)

These volumetric indexes can be used for the study of pathologies such like Parkinson disease, schizophrenia, Alzheimer disease, or longitudinal studies depending on age, for example.

2. Binding indexes

Binding indexes are computed from the corresponding registered functional acquisition as the MRI-segmented mask allows an accurate location of the ROI (Fig.10)

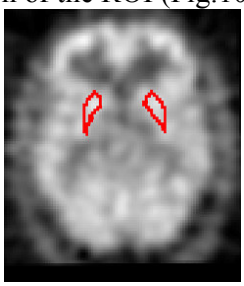


Fig. 10. Location of the regions of interest in the SPECT study

The computation of binding indexes is tracer-specific. The simplest indexes used are minimum, maximum and mean binding in the region of interest [ELFA-01]. These measures can't be compared for inter patients studies because tracer's elimination and binding are very dependent on patient's metabolism and on the moment of the acquisition after tracer injection. We thus compute an absolute measure, normalizing tracer binding in the specific region of interest by the binding obtained in non specific structures (structures in which the tracer isn't supposed to fix). Several normalized indexes are computed, among which normalized binding: R=(Avgspecif-Avgnonspecif)/Avgnonspecif, binding asymmetry, …

Depending on the tracer, the most commonly non specific reference structures used are: the cerebellum, parietal cortex, occipital cortex, frontal cortex (some authors also use the whole brain so that we propose it). Each reference has to be chosen according to the specificities of the tracer by the user.

On the whole, about ten indexes, normalized or not, are computed for each couple of structure, and for each structure independently.

The quantification step has also been assessed, first on simulated data and on real acquisitions; results were compared with the diagnosis manually obtained by an expert.

C. Conclusion

The results and methods for the whole process have already been published [FREN-03]. This quantification process has the advantage of being totally automatic and faster than manual tracing (less than one minute for the segmentation of a couple of structures, few seconds for indexes computation) while being as precise as the manual method. Results obtained are comparable with the manually obtained ones and clinical conclusions about the patients also. This process has also the advantage to allow a more accurate definition of the reference region for normalized binding indexes.
The originality of the procedure resides in the use of both acquisitions (MRI and SPECT) for quantification, allowing the patient to be his own anatomical reference. The use of possibilistic theory for each piece of information representation and the fusion of two kinds of information has also the advantage of allowing "abnormalities" in patient's morphology to be taken into account (which isn't possible with a binary shape reference, for example). This process is also fully automatic, but allows the user to check each step of the method and correct it if wanted.
There are many applications for this process, e.g. the extraction of the most pertinent set of indexes to evaluate a given pathology (for example using factorial analysis) and the possibility to obtain a pre-diagnosis using e.g. discriminant analysis.
Main application for this process is the study and pre-diagnosis of all neurotransmission-implied pathologies, with the possibility of differential diagnosis between Parkinson's disease and parkinsonian syndrome with a set of well-chosen indexes and a data base of healthy and pathological subjects.

References

[BARI-94] Barillot C et al. Data fusion in medical imaging: merging multimodal and multipatient images, identification of structures and 3D display aspects. Yearbook of Medical Informatics, 290-295, 1994.

[BARR-00] Barra V. in Fusion d'images 3D du cerveau : études de modèles et applications. Ph.D. Thesis, Université d'Auvergne, Clermont-Ferrand (France), 2000.

[BLOC-95] Bloch I and Maitre H. Fuzzy mathematical morphologies: a comparative study. Pattern recognition, 28:1341-1387, 1995.

[CATA-01] Catafau AM. Brain SPECT in clinical practice. Part I : Perfusion. In The Journal of Nuclear Medicine, 42:259-271, 2001.

[COLI-99] Colin A and Boire JY. MRI-SPECT fusion for the synthesis of high resolution 3D functional brain images: a preliminary study. Comput. Meth. Programs Biomed., 60:107-116, 1999.

[DHAW-96] Dhawan AP. et al. A system for MR brain image segmentation. Proceedings of the 18th IEEE/EMBS, paper 189, 1996.

[DUBO-88] Dubois D and Prade H. in Possibility Theory, an approach to the computerized processing of the uncertainty. Plenum Press; 1988.

[ELFA-01] El Fakrhi G et al. Absolute activity quantitation in simultaneous 123I/99mTc brain SPECT. In The Journal of Nuclear Medicine, volume 42, pages 300-308, 2001.

[FREN-01] Frenoux E et al. Segmentation of the striatum using data fusion. to appear in the Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Istanbul, Turkey, October 2001.

[FREN-03] Frenoux E et al, A new method for the quantitative study of neurotransmission, 25th IEEE Engineering in Medicine and Biology Society, Cancun, Mexico, september 2003 (to appear).

[GERA-98] Géraud T. in Segmentation des structure internes du cerveau. Ph. D. Thesis, Ecole Nationale Supérieure des Télécommunications, Paris (France), 1998.

[HABR-99] Habraken JBA et al. Quantification and Visualization of Defects of the Functional Dopaminergic System Using an Automatic Algorithm. In The Journal of Nuclear Medicine, volume 40, pages 1091-1097, 1999.

[HEIN-92] Heinz Hohne K and Hanson WA. Interactive 3D segmentation of MRI and CT volumes using morphological operations. J Comput Assist Tomogr, 16(2):285-294, 1992.

[HILT-01] Hiltner J et al. An approach to use linguistic and model-based fuzzy expert knowledge for the analysis of MRT images. Image and Vision Computing, 19:195-206, 2001.

[JANG-97] Jang DP et al. Contour detection of hippocampus using dynamic contour model and region growing. Proceedings of the 19th International IEEE/EMBS Conference Chicago (IL, USA), 1997.

[OZKA-94] Ozkan M et al. Neural-network-based segmentation of multimodal medical images: a comparative and prospective study. Yearbook of medical informatics, 302-312, 1994.

[SCHU-99] Schulz JB et al. Magnetic resonance imaging-based volumetry differentiates idiopathic Parkinson's syndrome from multiple system atrophy and progressive supranuclear palsy. In the Annals of Neurology, 45:65-74, 1999.

[TOUR-00] Touraille E and Boire JY, Elastic registration of MRI scans using fast DCT, World Congress on Medical Physics and Biomedical Engineering, Chicago, July 2000.

[VINI-95] Vinitski S. et al. 3D segmentation in MRI of brain tumors: preliminary results. Proceedings of the 17th IEEE/EMBS Conference, paper 192, 1995.

[ZADE-78] Zadeh L. Fuzzy sets as a basis for theory of possibility. International Journal of Fuzzy Sets and Systems, 1:3-28, 1978.

[ZHU-95] Zhu H et al. A deformable region model for locating the boundary of brain tumor. Proceedings of the 17th IEEE/EMBS Conference, paper 547, 1995.

Related publications in the ERIM research team

COLIN A AND BOIRE JY. A novel tool for rapid prototyping and development of simple 3D medical image processing applications on PCs, Comput. Meth. Prog. Biomed., 53:87-92, 1997.

COLIN A AND BOIRE JY. MRI-SPECT fusion for the synthesis of high resolution 3D functional brain images: a preliminary study in application to the medical follow up of a brain pathology, Comput. Meth. Prog. Biomed., 60 : 107-116, 1999.

BARRA V, BOIRE JY. Tissue characterization on MR images by possibilistic clustering on a 3D wavelet representation, JMRI, 11:267-278, 2000.

BARRA V, BOIRE JY. Automatic segmentation of subcortical brain structures in MR images using information fusion, IEEE Trans. Med., 20(7):549-558, 2001.

BARRA V, BRIANDET P, BOIRE JY. Fusion in medical imaging: theory, interests and industrial applications, MEDINFO 10:896-900, 2001.

BARRA V, FRENOUX E, BOIRE JY. Automatic volumetric measurement of lateral ventricles on MR images with correction of partial volume effects, JMRI, 15:16-22, 2002.

BARRA V, BOIRE JY. Automatic segmentation of subcortical brain structures in MR images using information fusion. In IEEE Transactions on Medical Imaging, volume 20, issue 7, pages 549-558, 2001.

FRENOUX E, BARRA V, BOIRE JY. Segmentation du striatum par fusion d'informations numériques et symboliques, Coll. Informatique et Santé, 13:173-180, ED. SPRINGER, 2002.

FRENOUX E, BARRA V, BOIRE JY. Segmentation of the striatum using data fusion, 23rd Annual Conference of the IEEE Engineering in Medicine and Biology, Istanbul, Turkey, October 2001.

FRENOUX E, BARRA V, BOIRE JY. Quantification of neurotransmission defects in functional imaging using information fusion: a prospective study, 9th International conference on Information Processing and Management of Uncertainty in knowledge-based systems, July 2002.

FRENOUX E, BARRA V, BOIRE JY, HABERT MO. A new method for the quantitative study of neurotransmission, 25th IEEE Engineering in Medicine and Biology Society, Cancun, Mexico, september 2003 (to appear).