Segmentation d'image de biopuces par Champ de Markov tolérant les déformations locales de grille.

Christophe Gouinaud

LIMOS-ISIMA, Campus des Cézeaux, 63177 Aubière, gouinaud@isima.fr

Résumé Notre travail a pour but de fournir une méthode de segmentation d'image permettant d'augmenter la qualité des mesures faites grâce à des biopuces. Nous présentons toutes les étapes nécessaires à la réalisation d'une segmentation et proposons diverses améliorations tout en les comparant aux méthodes usuelles. Enfin, nous montrons des résultats sur deux jeux d'images présentant des situation différentes.

1 Introduction

L'intérêt de l'utilisation des biopuces cdna pour la génétique n'est plus à démontrer [7]. Cette technologie complexe arrive maintenant à maturité et son utilisation s'étend à la modélisation des relations gène-expression-individu. De ce fait, le défi actuel est l'amélioration de la précision des mesures réalisées de façon à augmenter la qualité des expressions estimées et donc les résultats fonctionnels.

D'un point de vue concret, l'utilisation des biopuces fait appel à des technologies très différentes. Les sources d'imprécision dans le process de production sont nombreuses et il est donc fondamental de maîtriser chaque étape et d'appliquer avec rigueur chaque procédure. Parmi toutes les étapes, celle du traitement d'images est certainement celle où les gains de précision les plus importants peuvent être obtenus [4] [11]. Cette phase est évidemment affectée par toutes les étapes précédentes. Le travail présenté ici s'attaque donc à l'amélioration du calcul des expressions en proposant une nouvelle méthode de segmentation des images.

La première motivation de ce travail est de construire une méthode de traitement d'images plus générique que celle proposée par les concepteurs de robot spoteur et de scanner à biopuces. En effet, ces derniers sont souvent associés à des logiciels de traitement d'images développés dans le but d'utiliser les métadonnées décrivant la lame fournie par les robots d'un même constructeur [20]. Ceci complique en particulier les comparatifs entre des biopuces produites avec des appareils différents et donc les échanges de données images entre laboratoires.

Ce travail vise aussi à arriver à une qualité plus importante de segmentation des images et en particulier à gérer correctement les déformations de spots, à corriger au mieux les déplacements de sonde et donc à fournir les meilleurs masques possibles pour le calcul des expressions. La méthode proposée ici est originale car elle combine des approches différentes pour chaque étape nécessaire à la détection des spots. Elle s'inspire de méthodes utilisées dans d'autres contextes tels que la télédétection sur images SAR qui sont également acquises en lumière cohérente et présentent donc le même genre de

bruit [8]. Elle présente aussi l'intérêt de faire intervenir au minimum un opérateur tout en tolérant l'utilisation d'images issues de matériel quelconque.

Ce document est présenté en trois parties : la première est consacrée au contexte de cette étude, la seconde traite du problème de positionnement de la grille sur l'image et la dernière présente notre méthode de segmentation et ses résultats.

2 Contexte

Le calcul de l'expression des spots nécessite deux prètraitement, un de localisation des zones où se trouvent les sondes sur l'image, nous l'appellerons positionnement de la grille, et une réalisant le calcul de la liste des pixels appartenant à un spot qui est la segmentation. Le positionnement est essentiel pour éviter des erreurs d'association risquant de conduire à des résultats erronés d'expérience [19]. Il est couramment réalisé suivant deux méthodes qui sont l'utilisation de métadonnées ou l'association zone-spot après segmentation de l'image.

- l'utilisation de métadonnées nécessite l'emploi d'appareils et de logiciels d'un unique constructeur ou au minimum une interopérabilité entre tous les acteurs de la chaîne.
- la stratégie inverse de la précédente méthode consiste à réaliser la segmentation de l'image en spot puis à plaquer dessus la grille des sondes fournie par le spoteur [14]. Le but de cette méthode est de corriger les éventuels effets de glissement des spots ou autres artefacts introduits durant la production de la lame, la phase d'hybridation et l'acquisition de l'image. Il s'agit ici soit de processus d'association individuelle de spots, soit de méthode associant recherche des blocs de dépôt et association des spots dans les blocs. Après l'essai de diverses variantes de cette classe de méthodes [10], nous devons constater qu'il s'agit d'un problème ouvert car ces techniques ne peuvent s'affranchir des risques inhérents au glissement de sonde.

Face à ce constat, nous proposons ici une méthode semi-automatique de positionnement, décrite dans la partie 3, permettant de corriger les défauts de la deuxième classe de méthode de positionnement sans métadonnées.

Le but de la segmentation est de séparer les pixels d'un spot de ceux du fond. Les méthodes les plus couramment employées fonctionnent soit à l'aide d'algorithmes s'appuyant sur la reconnaissance des formes [9] ou bien à l'aide de classifications basées sur la valeur des pixels [21].

- dans le cas où on utilise des critères basés sur la forme, on présuppose que les spots visibles sur l'image ont des caractéristiques géométriques conformes aux sondes déposées. Les caractéristiques présupposés sont donc le plus souvent la circularité et la convexité du spot. Les techniques employées utilisent soit des cercles fixes ou ellipses adaptatifs mais le positionnement et la forme sont présupposés [17]. Bien évidemment, ces conditions ne sont que rarement réalisées et les résultats produits conduisent à l'intégration de nombreux pixels de fond dans les expressions calculées.
- l'utilisation de classification est basée sur l'existence d'une différence statistique entre les valeurs les pixels du fond et ceux dles spots. Ces méthodes sont le plus

souvent basées sur l'analyse de statistiques du premier ordre discriminant localement les pixels du spot de ceux du fond [14]. Elles sont parfois associées à des techniques de croissance de région ou de contours actifs qui réintroduisent l'aspect reconnaissance de forme.

Nous proposons ici une méthode hybride basée sur une double prise en compte de la forme et des signaux visant à s'affranchir des petites erreurs de positionnement et intégrant des hypothèses de connexité par arcdes spots. D'autres travaux ont développé des approches similaires telles que celles basées sur les classifications de Man-Whitney ou basées sur des techniques d'association dissociation ([3]). Notre approche se distingue par l'emploi de critères topologiques de voisinage associé une approche statistique évoluée. La partie **??** expose notre approche basée sur une segmentation markovienne.

3 Positionnement de la grille.

La première des tâches nécessaire à la segmentation de biopuces est le repositionnement de la grille des spots sur l'image scannée [13]. Ce sujet a fait l'objet de nombreux travaux, utilisant des grilles déformables, basés sur une prédétection des spots ou basés sur la reconnaissance de blocs de spots [1]. Cependant, les résultats de ces travaux se confrontent difficilement à la réalité des laboratoires et ont du mal à positionner correctement les grilles.

La double incertitude induite par les spots non exprimés et le positionnement incertain de la lame implique l'utilisation d'un positionnement semi-automatique. Le positionnement global de la grille se fait donc en désignant des pixels de l'image et leurs positions idéales et en utilsant un modèle de déformation.

Le choix du modèle doit être fait en fonction de la physique du système d'acquisition de façon à compenser les déformations par rapport à la prise de vue idéale. Dans le domaine des scanners à objectif confocal pour biopuces, la prise de vue est correcte quand ll'orientation de la lame est telle que son bord le plus long est parallèle à la trajectoire du centre de rotation du bras du scanner. Les déformations globales de l'image sont donc modélisables sous forme d'une combinaison d'applications affines en 2D dans le plan de l'image qui s'écrit :

$$\begin{cases} X^{'} = a_{11}X + a_{21}Y + a_{31} \\ Y^{'} = a_{12}Y + a_{22}Y + a_{32} \end{cases}$$

où X, Y sont les coordonnées d'un spot dans le repère de la grille, $X^{'}, Y^{'}$ dans le repère de l'image et les a_{ij} les coefficients du modèle.

Afin de limiter les erreurs, nous utilisons une méthode de détermination des paramètres par moindres carrés utilisant au minimum quatre points, ce qui permet d'absorber des petites erreurs de positionnement tout en distinguant les incohérences flagrantes. Pratiquement, nous affichons l'image et la grille placée sur l'image par un simple centrage puis l'opérateur indique au moins quatre amers en cliquant sur l'image et sur la grille. Ensuite, nous calculons le modèle d'interpolation et déplaçons la grille.

Après le calage global de la grille, nous constatons souvent des erreurs résiduelles de valeur suffisante qui empêche d'utiliser le bord de la zone comme étant certainement du fond.

L'observation des déplacements [2] nous conduit à distinguer le cas où le spot n'est pas à la place prévue mais sans être déformé, ou il est donc quasi circulaire, de celui ou il est déformé

Nous avons donc développé ici une approche duale basée sur la corrélation avec un modèle circulaire associé à un calcul barycentrique basé sur les luminances. Le premier algorithme vise à se recaler sur les spots qui n'ont que peu de déformations alors que le deuxième vise à prendre en charge les spots déformés et non uniformément exprimés.

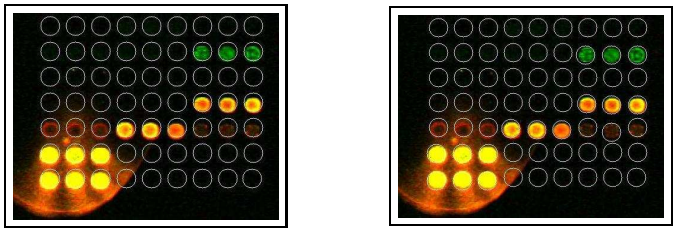

FIGURE 1. Affinage de la postion des spots, en haut avant en bas après.

Pratiquement, nous calculons la corrélation entre le premier canal image et un modèle circulaire binaire de même taille que les spots recherchés. Nous déplaçons ce modèle sur la fenêtre entourant le spot et calculons le corrélogramme la valeur maximum obtenue. Si celle-ci est supérieure à 0,75, nous considérons le point obtenu comme fiable. Nous réalisons ensuite le même calcul sur le deuxième canal. Nous calculons ensuite le barycentre de la fenêtre entourant le centre du spot donné par la grille nous disposons ainsi de positions différentes que nous fusionnons par un modèle linéaire dont les coéficients sont liées au valeurs de corrélation obtenu.

Cette technique s'est avérée particulièrement efficace dans le cas de spots ayant la forme de donuts et produit une grille affinée telle qu'elle est représentée sur la figure 1.

4 Segmentation

Comme cela est dit dans la partie précédente, notre méthode de segmentation nécessite, pour son initialisation, une estimation de la moyenne du spot et du fond l'entourant. Cette estimation est problématique parce qu'on ne connaît pas les pixels constituant le spot et que cette connaissance est notre but.

Pour estimer cette moyenne, on utilise classiquement les pixels entourant le centre du spot donné par la grille ou une forme circulaire représentant la forme idéale du spot. Ces deux approches ont pour défaut d'être trop sensibles au positionnement de la grille et d'augmenter les risques de confusion entre le fond et le spot. La méthode des centiles décrite dans [5] améliore un peu les estimations de moyenne mais reste trop conditionnée par la qualité géométrique des spots.

En observant un grand lot d'images et en réalisant de nombreuses segmentations manuelles, nous nous sommes aperçus que si la forme des spots pouvait être très variable, leur surface varie peu. En fait, les déformations de spot sont la plupart du temps

dues à un glissement de produit sur la lame. Notre méthode d'estimation est donc basée sur ce fait.

Nous procédons à l'estimation de la moyenne du spot en observant son influence sur la moyenne du fond entourant le spot. Pratiquement, cela consiste à écrire que la moyenne d'une zone (Z) contenant un spot est la somme pondérée de la moyenne des valeurs du fond (F) et des pixels (S) ce qui nous permet de déduire la moyenne $E(S)$ de la zone S en fonction de $E(Z)$ et lde $E(F)$:

$$E(S) = \frac{E(Z)card(Z) - E(F)card(F)}{card(S)}$$

La moyenne ($E(F)$) du fond est mesurée sur le bord de la boîte entourant le spot. La surface théorique du spot ($card(S)$) est donnée par les caractéristiques de la lame. La population du fond ($card(F)$) est donnée par $card(Z) = card(F) + card(S)$.

Avant même parler de notre segmentation, il faut poser les conditions d'emploi des deux canaux à notre disposition. Nous avons choisi de combiné les informations produite par les deux canaux car les biopuces sont conçus pour que l'information donnée par les deux images soit différente dans leur signification biologique. En théorie, on doit obtenir une même forme mais avec des intensités très différentes. Ces intensités peuvent être très proches ou très corrélées mais ne le sont pas dans le cas où le niveau d'expression d'un seul canal est très faible. Quand un spot s'exprime, cela signifie que la lumière perçue par les capteurs provient de la fluorescence et que la réponse du fond n'est plus visible. A l'inverse, dans une zone non hybridée, la lumière mesurée par le scanner provient uniquement du fond. Dans ce dernier cas, la sonde est transparente et ne fournit alors aucune information de forme exploitable et cela même si une certaine déformation du signal de fond peut être mesurée. Il convient donc d'adopter le schéma de fusion donnée suivant :

- Dans le cas numéro un, la zone segmentée comme appartenant à la sonde doit provenir uniquement de l'image bien exprimée,
- Dans le cas numéro deux, la zone d'expression est celle où les deux segmentations ont conclu à la classe spot,
- Dans le cas numéro trois, l'information du canal qui a une forme très différente du dépôt doit être rejetée ou sursegmentée vers celle de l'image bien exprimée. Il faut noter que ce dernier cas est très rare.

Ceci étant posé, nous avons choisi de segmenter séparément les images des deux canaux CY5 et CY3 puis d'appliquer le schéma de fusion défini ci-dessus en distinguant les cas par une post-estimation des moyennes. Concrètement, nous calculons les moyennes du fond et du spot, après segmentation sur chaque canal, dans l'image d'un canal en utilisant la forme segmentée sur l'autre canal et comparons les variations de statistiques sur l'un et l'autre. Ceci nous permet d'orienter notre fusion.

Il nous faut également parler des raisons du choix d'une classification Markovienne. Notre segmentation est basée sur le fait que les statistiques du premier ordre des spots sont différentes de celles du fond. Les biologistes sont intéressés par la médiane des pixels constituant les spots car ils considèrent celle-ci comme plus représentative de l'expression des gènes que la moyenne. Notre segmentation est cependant basée sur la moyenne pour au moins deux raisons :

– si nous segmentons grâce à la médiane, nous risquons de biaiser la mesure finale
– la moyenne est beaucoup plus rapide à calculer que la médiane qui demande de plus une réévaluation complète quand on ajoute ou retire un individu de l'espace d'étude. Cette rapidité est essentielle pour pouvoir exploiter notre travail dans un logiciel interactif tournant sur des ordinateurs personnels.

Nous utilisons donc une classification itérative basée sur la distance d'un pixel à la moyenne de la classe à laquelle il est censé appartenir. Dans ce cadre, une première approche très classique [16] qui s'appelle une classification par nuée dynamique peut être tentée. Concrètement, nous calculons dans ce cas pour chaque pixel la distance à la moyenne de chaque classe et en déduisons son appartenance. La distance utilisée est alors de la forme :

$$D(i, c) = |X - E(X_c)|$$

où $E(X_c)$ est la moyenne de la classe des pixels de la classe C. Nous appelerons D attache aux données dans la suite.

Une fois tous les pixels classés, nous réévaluons la moyenne des classes et reclassons chaque pixel. Ce procédé est réalisé localement pour chaque sonde sur une petite fenêtre entourant le spot. En effet, les valeurs d'expression étant très disparates d'un spot à l'autre, leur moyenne est très différente et donc une approche globale à l'image serait aberrante. Le résultat que nous obtenons est une image de masques, superposable à l'image d'origine, dont les pixels ont une valeur zéro pour le fond et une valeur correspondant au numéro du spot.

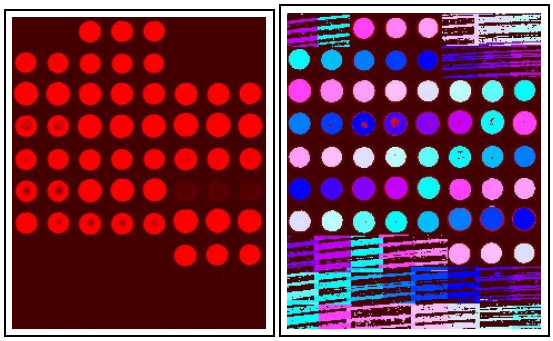

FIGURE 2. Classification par nuées dynamiques : à gauche, le canal rouge d'une image de biopuce, à droite, le résultat.

Sur la figure 2 nous montrons le résultat de ce traitement et nous pouvons constater qu'il est loin d'être parfait. Nous remarquons en particulier que certains spots, au centre de l'image, comportent des trous et que le traitement des zones de faible intensité aboutit à un résultat aberrant.

Les classifications telles que décrites ci-dessus n'utilisent qu'un critère basé sur la valeur des pixels. Ceci explique pourquoi beaucoup de pixels isolés sont mal classés comme on peut le voir sur la figure 2. Le problème vient du fait qu'il existe dans l'histogramme de l'image une zone confondue où se mélangent des pixels appartenant au fond et aux spots. Cette confusion ne peut être compensée que par l'introduction d'informations supplémentaires.

De ce fait, nous allons introduire dans notre classification des critères spatiaux qui nous permettront de mieux classer les pixels isolés ayant une valeur aberrante. Les spots étant le plus souvent connexes par arc, il apparaît qu'un pixel a forcément de nombreux voisins de la même classe que lui. Il est donc possible d'enrichir l'information de distance à l'aide du comptage des voisins de même classe. Nous avons donc modifié le calcul de la fonction d'appartenance à une classe de façon à relâcher l'attache aux données au profit des critères spatiaux. Ceci revient à modifier le calcul de la distance des classes pour mieux gérer la partie incertaine.

Dans la suite, nous allons exprimer le facteur d'appartenance à une classe comme un potentiel qui se manipule plus facilement qu'une probabilité. Celui-ci comporte deux termes, un exprimant l'appartenance à une classe et l'autre l'appartenance à l'autre classe. Soit $I\left(x,y\right)$ l'intensité du pixel de position x,y, nous écrivons le potentiel d'appartenir à la classe c relativement à la classe c' :

$$P(x,y,c,c') = \alpha\left(I\left(x,y\right),c\right) - \alpha\left(I\left(x,y\right),c'\right) + e\left(x,y,c\right) - e\left(x,y,c'\right)$$

où $\alpha\left(I,c\right)$ représente le potentiel d'appartenance à la classe c pour l'intensité I et $e(x,y,c)$ l'énergie potentielle fournie par le voisinage. La difficulté de mise au point d'une telle classification réside dans le choix de ces termes et dans leur équilibrage qui est toujours délicat. Pour la classe "fond" les contraintes sont :

- la nature du bruit fond, composé de chatoiement et de fragments de sonde hybridés et la connexité des spots, qui induit qu'ils sont sous-représentés en surface, nous a conduit à choisir des potentiels symétriques. En effet, nos objectifs pour la classe fond sont d'éliminer des pixels isolés de fortes valeurs classés à tort comme appartenant à un spot ce qui suppose d'utiliser un critère spatial fort.
- nous souhaitons que, pour des zones de moyenne proche de celle du fond mais incluse dans le spot, la classification les considère comme appartenant aux spots. Ceci implique une attache aux données forte autour de la moyenne du fond mais qui décroît rapidement.

Pour la classe "spot", les contraintes sont :

- les valeurs proches de la moyenne doivent être affectées à la classe "spot" sous réserve de voisinage.
- les pixels proches de la moyenne du fond doivent être rejetés sauf en frontière de façon à corriger les effets de sur-représentation surfacique.
- l'attache aux données doit croître au fur et à mesure qu'on s'éloigne de la moyenne du fond et que l'on s'approche de la moyenne du spot.

À partir de ces deux jeux de contraintes, nous avons fabriqué le potentiel linéaire par morceaux $\alpha\left(I,c\right)$ défini par les formules suivantes :

si $c = F$

$$\begin{cases} I < \overline{F} & \alpha(I,F) = 1 \\ \overline{F} < I < \frac{\overline{F}+\overline{S}}{2} & \alpha(I,F) = \frac{3}{8}\left(\frac{\frac{\overline{F}+\overline{S}}{2} - I}{\overline{S}-\overline{F}}\right) + \frac{1}{4} \\ \frac{\overline{F}+\overline{S}}{2} < I & \alpha(I,F) = 0 \end{cases}$$

si $c = S$

$$\begin{cases} I < \overline{F} & \alpha(I,S) = 0 \\ \overline{F} < I < \frac{\overline{F}+\overline{S}}{2} & \alpha(I,S) = \frac{1}{4} \\ \frac{\overline{F}+\overline{S}}{2} < I < \overline{S} & \alpha(I,F) = \frac{3}{8}\left(\frac{I - \frac{\overline{F}+\overline{S}}{2}}{\overline{S}-\overline{F}}\right) + \frac{1}{4} \\ \overline{S} < I & \alpha(I,F) = 1 \end{cases}$$

où $\overline{S}$ et $\_overline F$ sont les moyennes estimées du spot et du fond. Ce potentiel α est représenté graphiquement en fonction de l'intensité I sur la figure 3.

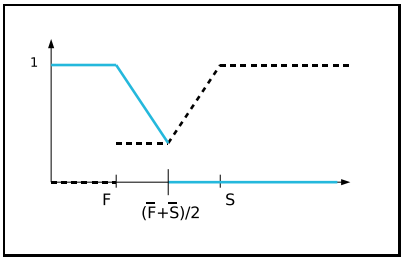

FIGURE 3. Sur cette figure nous voyons le potentiel α pour la classe "fond" en clair et pour la classe "spot" en tireté.

Le terme $e(x, y, c)$ est obtenu en comptant le nombre de voisins de la classe c se trouvant dans un voisinage en huit connexités, c'est-à-dire limité aux huit pixels entourant le spot. Ce choix est gouverné par le fait que les spots n'ont pas de direction privilégier et sont de taille réduite ce qui nous conduit à limité les effets à longue distance du processus de régularisation spatiale. Avec cet ensemble de contraintes, il faut maintenant fabriquer l'image de classe c minimisant la fonctionnelle P en tout point de l'image originale. Pour celà, certaines propriétés de notre fonctionnelle sont précieuses et notamment le fait que le processus de régularisation spatiale en huit connexités nous permet d'affirmer que nous sommes dans un cadre markovien [6]. Cela nous indique que le problème d'optimisation a une solution et que par conséquent, un schéma itératif de recherche par minimisations successives convergera vers cette solution.Cela implique également qu'il est possible de segmenter l'image spot par spot sans changer le résultat. Nous pouvons donc découper arbitrairement en morceaux l'image sans changer l'aspect de la solution.

Le problème principal que pose, par contre, un tel schéma est sa complexité. Une recherche directe du minimum absolu serait beaucoup trop longue, en particulier dans le cadre d'un logiciel interactif. Nous procèderons donc ici par recuit simulé.

Notre algorithme se déroule donc en trois phases : une phase d'initialisation, une phase de classification itérée et une phase de nettoyage éventuel du résultat.

La phase d'initialisation consiste à découper l'image en sous-imagettes autour des spots et estimer par la méthode décrite dans la partie 4 la moyenne du fond et du spot pour chaque imagette. Ensuite pour chaque site, on tire un masque de classe suivant un processus Poissonnien.

La phase de classification consiste pour chaque imagette à prendre un pixel et calculer le potentiel pour la classe courante donnée par l'image de classe. Si ce potentiel est négatif, on change la classe, sinon on calcule l'expression $x = log(\frac{P}{t})$ et on tire aléatoirement un nombre x' suivant une loi uniforme entre 0 et 1 et si $x' > x$, nous changeons la classe du pixel bien que le potentiel indique que le pixel est plutôt bien classé. Ce choix permet de ressortir des minima locaux de potentiel sans devoir essayer toutes les combinaisons possibles. Nous passons ensuite aux pixels suivants. Puis nous réitérons le procédé avec $t_n = 0.95 * t_{n+1}$. La classification est finie quand le nombre de pixels qui ont été modifiés est inférieur à 1 pour 1000. La phase de nettoyage, parfois appelée trempe, consiste à attribuer les éventuels pixels isolés à la classe qui les entoure par simple comptage.

Ces algorithmes demandent beaucoup de calculs mais les propriétés des champs de Markov nous permettent de paralléliser sans risque le calcul des différentes imagettes. Celà raccourcit significativement la durée du traitement sur les machines dual-core ou multiprocesseurs, et l'on obtient une classification en quelques secondes pour des images courantes.

La partie suivante présente nos résultats.

5 Résultats de la segmentation markovienne.

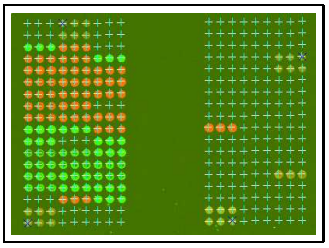

FIGURE 4. Image simple de biopuces.

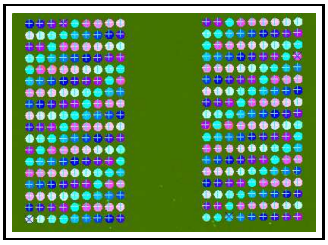

FIGURE 5. Image de la figure 4 segmentée à l'aide la technique du cercle fixe.

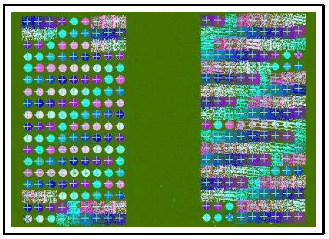

FIGURE 6. Image de la figure 4 segmentée à l'aide d'une classification par nuées dynamiques.

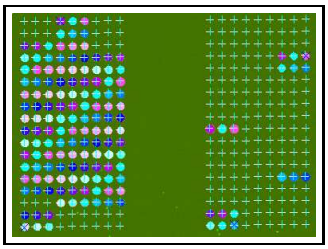

FIGURE 7. Image de la figure 4 segmentée à l'aide d'une classification par notre technique.

Nous ferons la présentation de nos résultats sur deux séries d'images, une série facile, telle celle de la figure 4, et une série plus difficile, telle celle de la figure 8, pour montrer les avantages et les limites de notre approche. Comme cela est dit précédemment, une bonne segmentation doit comporter au moins trois qualités : ne pas segmenter les spots non hybridés, se comporter correctement en présence d'artefacts, segmenter au plus juste les spots hybridés. Nous présentons nos résultats en référence à ceux obtenue par nuée dynamique et par les cercles fixes car se sont deux type de classification utilisé dans les outils professionnels dont dispose nos partenaires biologiste.

Pour juger de la première qualité, nous ne pouvons pas considérer directement le nombre de spots hybridés détectés. Il convient donc de calculer la probabilité de se tromper qu'a un algorithme quand il détecte un spot. Le taux de détection que nous obtenons est alors de 99 % des bons spots et 99,9 % des spots non hybridés ne sont pas segmentés. Ces résultats sont bien évidemment meilleurs que ceux des deux autres méthodes présentées.

Le comportement en présence d'artefacts ne peut se juger que visuellement car il n'existe pas de recensement exhaustif des défauts possibles dans ce type d'image. Les deux types les plus gênants sont liés au bruit de fond et aux arrachements de matière sur les sondes qui produisent des zones très brillantes après hybridation. Sur la figure 10, nous présentons le résultat de notre traitement dans le cas d'une image plus difficile (figure 8). Sur ces données, nous voyons plusieurs cas de pollution de sonde provenant d'arrachements d'autres spots avec différentes tailles et proximités de la sonde polluée. Nous constatons alors que le comportement de notre méthode est globalement satisfaisant, car les zones de pollution ne sont pas segmentées dès qu'elles atteignent une taille significative.

Pour ce qui est de juger de l'aptitude d'une segmentation à intégrer un minimum de pixels de fond dans les spots produits, la seule méthodologie valable consiste à tracer à l'aide d'un expert un grand nombre de spots à la main puis à calculer les matrices de confusion entre les segmentations et le masque manuel [12]. Cette opération est longue et fastidieuse si nous voulons traiter un grand nombre de cas. Nous avons donc procédé en tirant au sort 20 spots dans 32 lames différentes ce qui permet de balayer un grand nombre de cas sans dépenser trop de temps. Parmi ces spots, 15 par puce ont été

utilisés pour calculer la matrice de confusion et pour vérifier sa robustesse. Ces mesures montrent un gain de 15 à 18 % de bonne détection.

Notre méthode a donc les trois qualités qui correspondent à notre cahier des charges. Nous remarquons de plus que l'adaptation de notre segmentation aux petites déformations de grille, non compensées par les deux techniques de calages présentées précédemment, est excellente. Ceci montre bien la supériorité des méthodes à base de classification pour la détection de la forme des spots sur les images de biopuces.

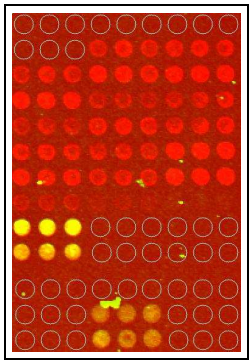

FIGURE 8. Image brut, les spots non remplit sont encerclés.

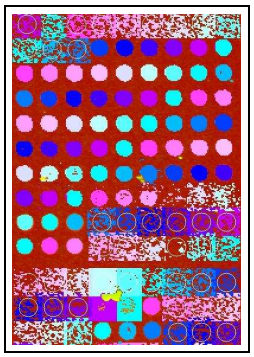

FIGURE 9. Résultat d'une classification par nuées dynamiques.

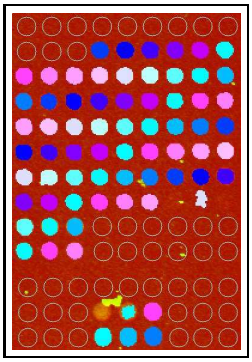

FIGURE 10. Résultat de notre traitement sur l'image de la figure 8.

6 Conclusions et perspectives

Le travail présenté ici atteint les buts que nous nous sommes fixés, car notre méthode permet de traiter des biopuces provenant de constructeurs quelconques tout en segmentant de façon robuste les forts et faibles contrastes. Notre méthode exploite au mieux les propriétés des différentes données disponibles (grille, images, ...) et elle utilise au minimum les interventions d'opérateurs.

Notre approche va être poursuivie suivant deux axes visant toujours l'amélioration de la qualité des mesures d'expression. Dans un autre aspect, nous travaillons sur les techniques de calibrage rouge vert afin d'améliorer cette fois la qualité de comparaison entre deux populations de gènes qui est souvent le but ultime de l'utilisation de biopuces.

Nous tenons à remercier le professeur Pierre Peyret de l'Université Blaise Pascal qui nous a fourni les images nécessaires à ce travail et Anne Moné qui à participé à l'expertise des résultat.

Références

1. G. Antoniol and M. Ceccarelli. Microarray image gridding with stochastic search based approaches. *Image and Vision Computing*, 25(2) :155 – 163, 2007.
2. Y. Balagurunathan, E. Dougherty, and Y. Chen. Simulation of cdna microarrays via a parameterized random signal mod el. *Joural of Biomediacl Optics*, 7(3) :507–523, July 2002.
3. V. Barra. Robust segmentation and analysis of dna microarray spots using an adaptative split and merge algorithm. *Computer Methods and Programs in Biomedicine*, 81(2) :174–180, Feb. 2006.
4. C. S. Brown, P. Goodwin, and P. K. Sorger. Image metrics in the statistical analysis of dna microarray data. *PNAS*, 96(16) :8944–8949, July 2001.
5. O. Demirkaya, M. H. Asyali, and M. M. Shoukri. Segmentation of cdna microarray spots using markov random field modeling. *Bioinformatics*, 21(13) :2994–3000, 2005.
6. X. Descombes. *Champs markoviens en analyse d'images*. 93 E 026, ENST, Paris-France, 1993.
7. M. B. Eisen and P. O. Brown. Dna arrays for analysis of gene expression. *Methodes in enzimology*, 303 :179–205, 1999.
8. J. W. Goodman. Some fundamental properties of speckle. *J. Opt. Soc. Am.*, 66(11) :1145–1150, 1976.
9. R. Hirata, J. Barrera, R. F. Hashimoto, D. O. Dantas, and G. H. Esteves. Segmentation of microarray images by mathematical. *Real-Time Imaging*, 8(6) :491–505, Dec. 2002.
10. M. Katzer and F. Kummert. A markov random field model of microarray gridding. In *in Proc. ACM Symposium on Applied Computing (SAC*, pages 72–77. ACM Press, 2003.
11. M. Katzer, F. Kummert, and G. Sagerer. Robust automatic microarray image analysis. In *In BREW Bioinformatics Research and Education Workshop*, 2002.
12. M. G. Kendall and A. Stuart. *The Advanced Theory Of Statistics*, volume 1. Charles Griffin and Compagny -Limited London and High Wycombe, third edition, 1952.
13. A. Kuklin, S. Shams, and S. Shah. Automation in microarray image analysis with autogene(tm). *Journal of the Association for Laboratory Automation*, 5(5) :67 – 70, 2000.
14. A. W.-C. Liew, H. Yan, and M. Yang. Robust adaptive spot segmentation of dna microarray images. *Pattern Recognition*, 36(5) :1251 – 1254, 2003.
15. M. McGoven and R. Fayek. Advantages of laser confocal microarray scanning. *Microarray Image Anlysis-Nuts and Bolts*, pages 51–68, 2002.
16. N. Pal and S. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26 :1277–1294, 1993.
17. A. Petrov and S. Shams. Microarray image processing and quality control. *Journal of VLSI Signal Processing Systems*, 38(3) :211–226, Nov. 2004.
18. A. Pretrov, S. Sha, S. Draghici, and S. Shams. Microarray image processing and quality control,. *Microarray Image Analysis-Nuts & Bolts*, (6) :99–130, 2002.
19. C. K. Wierling, M. Steinfath, T. Elge, S. Shculzen-Kremer, P. Aanstad, M. Clark, H. Lehrach, and R. Herwig. Simulation of dna hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC bioinformatics*, 3(29) :17, Oct. 2002.
20. Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cdna microarray data. *Journal of Computational and Graphical Statistics*, 11 :108–136, 2002.
21. C. Yidong, R. Edward, and all. Ratio-based decisons and the quantitative analysis of cDNA microrarray images. *Journal of biomedical optics*, 2(4) :364–374, Oct. 1997.