Prédiction du grade d'un cancer du sein par la découverte de motifs séquentiels contextuels dans des puces à ADN

Julien Rabatel[*,**], Mickaël Fabrègue [*], Sandra Bringay[*,***], Pascal Poncelet[*], Maguelonne Teisseire[****]

[*]LIRMM, Université Montpellier 2, CNRS
161 rue Ada, 34392 Montpellier Cedex 5, France
[**]Tecnalia, Cap Omega, Rond-point Benjamin Franklin - CS 39521
34960 Montpellier, France
[***]Dpt MIAP, Université Montpellier 3, Route de Mende
34199 Montpellier Cedex 5, France
[****]TETIS, 500 rue Jean-François Breton
34093 Montpellier Cedex 5, France

Résumé. Le cancer du sein reste de nos jours un problème de santé majeur et un véritable défi pour les biologistes et les professionnels de santé. Les puces à ADN permettent aujourd'hui d'étudier selon un jour nouveau les problématiques associées à cette maladie. Dans cet article, nous proposons de traiter les données issues des puces à ADN par le biais de l'extraction de motifs séquentiels contextuels (séquences de gènes ordonnés selon leur niveau d'expression associées à un contexte). L'objectif est de proposer une aide au diagnostic du grade d'une tumeur. Notre approche tient à la fois compte de l'information contenue dans les puces à ADN (exprimée par le biais de motifs séquentiels), mais également d'informations additionnelles d'ordre contextuel (e.g., âge du patient, taille de la tumeur, etc.) et qui sont associées aux données de puces à ADN lorsque celles-ci sont publiées en ligne. L'approche proposée a été évaluée sur des données réelles.

1 Introduction

De nos jours, le cancer du sein représente un problème de santé majeur. Selon le site `breastcancer.org`, environ 1 femme sur 8 aux Etats-Unis est susceptible de développer un cancer du sein au cours de sa vie, et on estime à plus de 200 000 le nombre de nouveaux cas pour l'année 2010. De nouveaux outils de diagnostic sont continuellement créés pour détecter des types spécifiques de cancer et de nouveaux traitements sont également développés pour s'appliquer à ces différents types de cancer. L'objectif est ainsi de réduire les effets indésirables tels que des disfonctionnements cardiaques ou des ménopauses prématurées par exemple en proposant aux patients des traitements ajustés aux spécificités de leur cancer. Cependant, malgré les nombreuses avancées dans le domaine, jusqu'à 50% des femmes touchées développeront des métastases distantes, qui restent malheureusement incurables.

Les trois principaux défis associés aux cancer du sein aujourd'hui sont les suivants : *(1)* comment diagnostiquer un cancer du sein le plus tôt possible (dépistage) et identifier le type d'une tumeur, *(2)* comment prédire la réaction d'un patient à un traitement donné en fonction des informations disponibles sur les historiques des précédents patients (sur leur type de tumeur et sur les résultats des traitements qu'ils ont reçu), et *(3)* comment proposer un choix de thérapie pour un patient donné en fonction des prédictions précédentes et en étant capable de l'informer sur ses chances de rémission, sur le développement prévisible de la maladie ainsi que les conséquences possible des traitements proposés.

Dans cet article, nous proposons d'exploiter les connaissances qui ont été rendues publiques par différentes équipes de biologistes travaillant sur le cancer du sein et qui sont disponibles sur Internet pour commencer à apporter des solutions aux trois défis précédents. Ces équipes ont mutualisé les résultats issus de l'analyse de puces à ADN appliquées sur les tissus de différents types de cancer du sein. Les puces à ADN sont des outils puissants permettant de dresser un véritable portrait génétique d'un échantillon biologique (ici des échantillons de tumeurs) en comparant l'expression de milliers de gènes dans différents tissus, cellules, ou conditions. Simon et Dobbin (2003) décrivent trois moyens d'exploiter les puces à ADN :

– **Comparer des classes** consiste à identifier des variations (e.g., dans l'expression des gènes) entre plusieurs classes. Il est possible de comparer des tissus normaux à des tumeurs (Alon et al. (1999)), ou des tumeurs qui réagissent bien à une thérapie à celles qui ne réagissent pas (Rosenwald et al. (2002)), ou encore de distinguer différents types de tumeurs (Dougherty (2001); Sotiriou et al. (2003)). Il est ainsi possible d'identifier des gènes ayant des comportements différents dans différentes classes et que l'on suppose jouer un rôle important dans une maladie. Ces informations peuvent alors être utilisées dans un but de prédiction.

– **Découvrir des classes** consiste à découvrir de nouvelles catégories dans une population (e.g., de nouvelles catégories de tumeurs). Par exemple, Sotiriou et al. (2003) a utilisé les données issues de l'analyse de puces à ADN appliquées à des tissus issus de tumeur du sein pour décrire différents types de cancer du sein.

– **La classification** exploite les résultats précédents pour affecter une nouvelle puce à ADN à une classe connue (e.g., pour associer cette nouvelle puce à un type de tumeur dans Van De Vijver et al. (2002)). Si le classifieur construit est suffisamment fiable, ses résultats peuvent être utilisés pour assister les professionnels de santé dans leur prise de décision lors d'un dépistage, par exemple.

Le nombre de données générées à partir de l'analyse de puces à ADN est considérable. Les méthodes statistiques et les méthodes de fouille de données jouent alors un rôle important dans la découverte de nouvelles connaissances. Cependant, leur mise en place est difficile du fait du grand nombre de valeurs mesurées par puce comparé au nombre de puces testées. Ce problème est connu sous le nom de « fléau de la dimension » (*curse of dimensionality*) (Dougherty (2001)). Par ailleurs, il existe des corrélations entre les expressions des gènes qui ne sont pas toujours bien connues et les valeurs d'expression des gènes sont également souvent entachées de bruit lié aux dispositifs expérimentaux. Pour toutes ces raisons, le problème de

A

Puce	G_1	G_2	G_3	G_4	G_5
$P1$	7.3	6.6	6.6	9.5	8.1
$P2$	5.6	7.4	5.6	5.3	7.9
$P3$	5.7	5.2	8.7	6.8	6.2

B

Puce	Séquence
$P1$	$\langle(G_2G_3)(G_1)(G_5)(G_4)\rangle$
$P2$	$\langle(G_4)(G_1G_3)(G_2)(G_5)\rangle$
$P3$	$\langle(G_2)(G_1)(G_5)(G_4)(G_3)\rangle$

FIG. 1 – **A** - *Puces à ADN.* **B** - *Séquences correspondantes.*

la classification de puces à ADN est différent des problèmes de classification connus et les méthodes traditionnelles rencontrent peu de succès (Zupan et al. (2000)).

L'objectif de l'étude menée ici est de développer une méthode pour extraire puis exploiter des séquences de gènes ordonnés selon leur niveau d'expression ainsi que des informations relatives au contexte lié à chaque patient (e.g., son âge) dans un but de classification, afin d'aider les professionnels de santé à diagnostiquer un type de cancer et à choisir la meilleure thérapie possible. Les séquences sont générées à partir des données issues de l'analyse de puces à ADN et sont basées sur une technique largement utilisée en fouille de données : l'extraction de motifs séquentiels. Salle et al. (2009) décrit un algorithme efficace pour extraire de telles séquences et montre l'intérêt de ces motifs pour distinguer différentes classes. Par exemple, le motif $\langle(17aag\_ovca\_dn)(tgz\_adip\_up)\rangle\ 80\% - Gr1,\ 40\% - Gr2$ signifie *« Pour 80% des tumeurs de Grade 1 et 40% des tumeurs de Grade 2, le niveau d'expression du gène* $17aag\_ovca\_dn$ *est inférieur à celui du gène* tgz_adip_up *»*. Les motifs séquentiels ont déjà été utilisés avec succès pour la classification de textes (Jaillet et al. (2006)). Nous allons montrer dans la suite leur pertinence pour la classification de tumeurs malgré la complexité des données et la teneur interdisciplinaire de ces travaux.

La suite de l'article est organisée de la manière suivante. Nous introduisons dans la section 2 l'extraction de motifs séquentiels contextuels dans les puces à ADN. La section 3 décrit la méthode de classification développée. L'évaluation de notre approche est exposée dans la section 4. Enfin, nous concluons et discutons les perspectives de ces travaux dans la section 5.

2 Extraction de motifs séquentiels dans les puces à ADN

La première étape de nos travaux concerne l'extraction des motifs séquentiels dans les puces à ADN. Dans un premier temps, nous décrivons comment les puces à ADN peuvent être traduites sous-forme d'une base de séquences, sur laquelle les sous-séquences fréquentes seront extraites. Ensuite, nous montrons comment les informations additionnelles liées à chaque patient (e.g., l'âge, la taille de la tumeur, etc.) peuvent être exploitées de manière à affiner nos connaissances.

2.1 Puces à ADN

La figure 1-A présente la structure des puces à ADN que nous manipulons. Soit G_1, G_2, ..., G_5 les gènes traités, et $P1, P2, P3$ des puces à ADN. Une puce contient une valeur d'expression pour chacun des gènes. Par exemple, la valeur de l'expression du gène G_1 dans la puce $P1$ est 7.3.

2.2 Motifs séquentiels et puces à ADN

Les motifs séquentiels traditionnels, introduits par Agrawal et Srikant (1995), peuvent être considérés comme une extension du concept d'itemsets fréquents de Agrawal et al. (1993) en considérant les estampilles temporelles associées aux items. La fouille de motifs séquentiels vise à l'origine à extraire des ensembles d'items fréquemment associés au cours du temps. En considérant l'étude des achats dans une boutique, un motif séquentiel pourrait par exemple être : «*40 % des clients achètent une télévision, puis plus tard achètent un lecteur DVD*».

Salle et al. (2009) propose une réprésentation des puces à ADN permettant l'extraction de motifs séquentiels dans le but de mieux caractériser différentes classes de puces. La découverte de motifs séquentiels dans les puces à ADN est définie comme suit.

Soit $\mathcal{G}$ un ensemble de ***gènes*** distincts. Un ***itemset*** est un sous-ensemble de gènes, noté $I = (g_1 g_2 ... g_n)$, tel que les gènes $g_1, ... g_n$ ont la même expression notée $exp(I)$. Une ***séquence*** s est une liste ordonnée d'itemsets notée $\langle I_1 I_2 ... I_k \rangle$, telle que $exp(I_1) < exp(I_2) < ... < exp(I_k)$. Ainsi, les gènes dans une séquence sont organisés en fonction de l'ordre de leur expression dans une puce donnée.

Exemple 1. *La figure 1-B montre, pour chaque puce, la séquence correspondante. Ainsi, la puce $P1$ est traduite par la séquence $\langle (G_2 G_3)(G_1)(G_5)(G_4) \rangle$.*

Soit $s = \langle I_1 I_2 ... I_m \rangle$ et $s' = \langle I'_1 I'_2 ... I'_n \rangle$ deux séquences. La séquence s est une ***sous-séquence*** de s', noté $s \sqsubseteq s'$, si $\exists i_1, i_2, ... i_m$ avec $1 \leq i_1 < i_2 < ... < i_m \leq n$ tel que $I_1 \subseteq I'_{i_1}$, $I_2 \subseteq I'_{i_2}$, ..., $I_m \subseteq I'_{i_m}$.

Exemple 2. *Soit les séquences $s = \langle (\mathbf{G_3})(\mathbf{G_4}) \rangle$ et $s' = \langle (G_2 \mathbf{G_3})(G_1)(G_5)(\mathbf{G_4}) \rangle$. s est une sous-séquence de s', i.e., $s \sqsubseteq s'$.*

Une ***base de séquences*** $\mathcal{B}$ est une relation $\mathcal{R}(ID, S)$, où un élément $id \in dom(ID)$ est un identifiant de séquence, et $dom(S)$ est l'ensemble des séquences. La ***taille*** de $\mathcal{B}$, notée $|\mathcal{B}|$, est le nombre de tuples dans $\mathcal{B}$. Un tuple $\prec id, s \succ$ ***supporte*** une séquence α si α est une sous-séquence de s, i.e., $\alpha \sqsubseteq s$. Le ***support*** d'une séquence α dans la base de séquences $\mathcal{B}$ est la proportion de tuples dans $\mathcal{B}$ supportant α, i.e. :

$$sup_{\mathcal{B}}(\alpha) = \frac{|\{\prec id, s \succ | \ (\prec id, s \succ \in \mathcal{B}) \wedge (\alpha \sqsubseteq s)\}|}{|\mathcal{B}|}.$$

Etant donné un nombre réel $minSup$ le seuil de ***support minimum***, tel que $0 < minSup \leq 1$, une séquence α est un ***motif séquentiel*** dans la base de séquences $\mathcal{B}$ si son support dans $\mathcal{B}$ est supérieur ou égal à $minSup$, i.e., $sup_{\mathcal{B}}(\alpha) \geq minSup$. La séquence α est alors dite ***fréquente dans*** $\mathcal{B}$.

Exemple 3. *Considérons la base de séquences $\mathcal{B}$ formée par les séquences de la figure 1-B, et un support minimum $minSup$ fixé à 1. Une séquence est un motif séquentiel si son support est supérieur ou égal à $minSup \cdot |\mathcal{B}| = 3$. La séquence $s = \langle (G_2)(G_5) \rangle$ est un motif séquentiel. En effet, son support est $sup_{\mathcal{B}}(s) = 3/3$ et est égal à $minSup$. En revanche, $s' = \langle (G_1)(G_4) \rangle$ n'est pas un motif séquentiel : son support est $sup_{\mathcal{B}}(s') = 2/3$ et est inférieur à $minSup$.*

2.3 Informations contextuelles

La découverte de motifs séquentiels dans les puces à ADN représentées sous forme de séquences permet d'extraire des connaissances précises sur différentes classes de données. Par exemple, il est possible de révéler que le motif $\langle(G_2)(G_5)\rangle$ est plus fréquemment associé aux cancers de grade 1 qu'aux autres grades. Ainsi, les motifs séquentiels extraits aideront à construire le profil génétique des différents grades de cancer, information utile pour effectuer une classification.

Cependant, les puces à ADN liées au cancer du sein sont fréquemment associées à des informations additionnelles. Par exemple, il sera possible de connaître l'âge du patient atteint, la taille de la tumeur détectée, etc. Ces informations, d'ordre contextuel, doivent être prises en compte afin de permettre une caractérisation plus fine des différents types de cancer, et ainsi permettre une classification plus fiable. Par exemple, nous cherchons à extraire des informations du type *« la séquence $\langle(G_3)(G_1G_4)\rangle$ est spécifique aux cancers de grade 1 chez les patients âgés de moins de 60 ans, tandis que la séquence $\langle(G_2)(G_3)\rangle$ est commune à tous les cancers de grade 2, peu importe l'âge du patient. »*

Rabatel et al. (2010) propose une définition formelle de telles informations contextuelles, ainsi qu'un algorithme d'extraction de motifs séquentiels contextuels (i.e., liés aux contextes auxquels ils sont spécifiques).

Les concepts liés à l'extraction de motifs séquentiels contextuels sont brièvement présentés dans l'exemple suivant.

Exemple 4. *Soit deux **dimensions contextuelles** Age (l'âge du patient lors du diagnostic) et Grade (le grade du cancer diagnostiqué). Nous noterons $dom(Age) = \{jeune, âgé\}$ le domaine de la dimension Âge et $dom(Grade) = \{1, 2\}$ le domaine de la dimension Grade. Un **contexte** est noté $[a, g]$ où $a \in dom(Age)$ et $g \in dom(Grade)$.*

Ainsi, le contexte $[jeune, 1]$ contient toutes les puces associées à un patient jeune atteint d'un cancer de grade 1. En introduisant une valeur joker $*$, nous pouvons également définir des contextes plus généraux. Par exemple, le contexte $[*, 1]$ correspond aux puces associées à un cancer de grade 1, pour un âge quelconque (i.e., $[*, 1]$ contient toutes les puces des contextes $[jeune, 1]$ et $[âgé, 1]$). Le contexte $[*, *]$ est quant à lui le contexte le plus général. Il représente l'ensemble des puces (i.e., pour un âge et un grade quelconque).

Les différents contextes définis peuvent ainsi être organisés selon la hiérarchie présentée dans la figure 2 (où j et a signifient respectivement $jeune$ et $âgé$).
*Dans cette hiérarchie, un motif séquentiel s est dit **spécifique à un contexte** c ssi :*

1. *il est fréquent dans c ainsi que dans tous les contextes descendants de c (s est alors dit **c-général**),*
2. *il n'existe pas de contexte plus général que c qui vérifie la première condition.*

Par exemple, le motif séquentiel s est spécifique à $[*, 1]$ ssi s est $[*, 1]$-général (i.e., il est fréquent dans $[jeune, 1]$ et $[âgé, 1]$) et s n'est pas $[*, *]$-général.

3 Classification de puces à ADN

Nous avons montré dans la section précédente comment extraire des motifs séquentiels contextuels dans des données issues de puces à ADN. Désormais, nous visons à exploiter ces

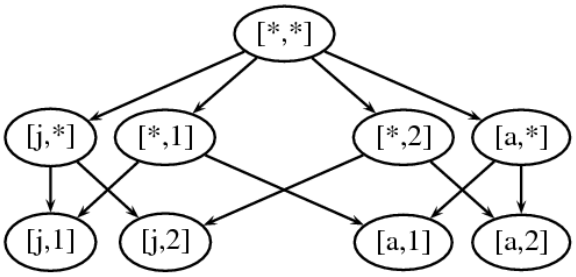

FIG. 2 – *Une hiérarchie de contextes.*

motifs pour améliorer la classification des puces. Le processus de classification de puces à ADN est constitué de trois étapes : *(i)* l'extraction des motifs séquentiels spécifiques pour chaque contexte considéré, *(ii)* la sélection des motifs les plus discriminants pour chaque contexte, et *(iii)* la prise de décision, i.e., l'association d'un grade à une puce à ADN donnée.

3.1 Définition du problème de classification

Dans cet article, le problème de classification que nous proposons de résoudre consiste en l'estimation du grade d'un cancer, à partir d'une puce à ADN liée à un patient. Cette méthode pourrait bien sûr s'appliquer pour d'autres objectifs de classification si l'on considère d'autres critères de classification des types de cancer que les grades comme proposé par Sotiriou et al. (2003). Pour ce faire, nous proposons d'utiliser les informations contextuelles disponibles afin de tirer profit de la précision des motifs séquentiels extraits.
Parmi les dimensions contextuelles définies dans la section 2, nous distinguons deux ensembles :

- **La dimension de classement.** Il s'agit de la dimension pour laquelle on souhaite faire une prédiction, désigner une valeur. Dans notre cas, la dimension *Grade* est la dimension de classement, i.e., nous ne connaissons pas sa valeur et nous voulons la prédire.
- **Les dimensions guides.** Il s'agit des dimensions dont nous connaissons la valeur, et sur lesquelles nous souhaitons nous appuyer pour mieux estimer la valeur des dimensions de classement. Par exemple, la dimension *Age* est une dimension guide dans le problème de classification des puces à ADN.

Soit $\mathcal{D}$ l'ensemble des dimensions contextuelles considérées. $D_1^g, ..., D_n^g$ sont les dimensions guides dans $\mathcal{D}$ et D^c est la dimension de classement dans $\mathcal{D}$.

Soit un contexte $c = [d_1^g, ..., d_n^g, d^c]$, où d_i^g est la valeur associée à la dimension D_i^g pour $i \in \{1, ..., n\}$ et d^c est la valeur associée à la dimension D^c. Le **guide** de c, noté $(d_1^g, ..., d_n^g)$, représente l'ensemble des valeurs associées à c sur les dimensions guides de $\mathcal{D}$.

Par la suite, nous nous intéresserons au cas où le guide d'un contexte est limité à une seule dimension. Ainsi, un contexte sera de la forme $[d^g, d^c]$.

Exemple 5. *Etant donné un patient âgé chez qui un cancer a été diagnostiqué, le problème de classification consiste à répondre à la question « Quel est le grade de son cancer (i.e., la valeur de la dimension de classement Grade), étant donné la puce à ADN de ce patient et son âge (i.e., la valeur de la dimension guide Age) ? »*

En d'autres termes, d'après la hiérarchie de contextes de la figure 2, il s'agira de décider, parmi les contextes $[\hat{a}g\acute{e}, 1]$ *et* $[\hat{a}g\acute{e}, 2]$*, quel est le plus proche du patient.*

3.2 Sélection des motifs séquentiels

Chaque contexte de la hiérarchie est associé à un ensemble de motifs séquentiels spécifiques. Cependant, tous ces motifs n'ont pas le même intérêt dans un but de classification. Ainsi, nous sélectionnons les k meilleurs motifs spécifiques pour chaque contexte.

3.2.1 Contextes conflictuels

L'utilisation d'une dimension guide permet de limiter le nombre de contextes possibles lors du processus de classification. Considérons l'exemple 5. L'âge du patient étant connu, il ne reste que deux contextes dans lesquels il peut être classé : $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$. En effet, les autres contextes ne respectent pas la contrainte d'âge que nous devons prendre en compte (e.g., $[*, 1]$ ou $[jeune, 2]$), ou ne contiennent pas de valeur pour la dimension de classement (e.g., $[\hat{a}g\acute{e}, *]$). Ces deux contextes sont conflictuels, i.e., le processus de classification doit être capable de les départager. Cette notion est importante car elle permettra par la suite de sélectionner les motifs qui seront les plus utiles pour la classification, i.e., qui seront les plus discriminants en considérant uniquement les contextes conflictuels.
De manière plus formelle, la notion de contextes conflictuels est définie comme suit.
Soit deux contextes $c = [d^g, d^c]$ et $c' = [d'^g, d'^c]$. Le contexte c' est **en conflit avec** c ssi :

- $d^g = d'^g$, i.e., les guides de c et c' sont identiques ;
- $d^c \neq *$ et $d'^c \neq *$, i.e., une valeur est bien définie sur la dimensions de classement ;
- $d^c \neq d'^c$, i.e., les deux contextes sont distincts.

Par la suite, nous noterons $conf(c)$ l'ensemble des contextes en conflit avec c.

3.2.2 Sélection

Soit un contexte c donné, nous cherchons l'ensemble des k motifs qui seront les plus pertinents dans un but de classification, i.e., les motifs les plus discriminants relativement aux autres contextes.

Nous définissons par conséquent une **mesure d'intérêt d'un motif** m pour caractériser un contexte c, notée $disc_c(m)$ et définie comme suit :

$$disc_c(m) = sup_c(m) - \max_{x \in conf(c)} sup_x(m).$$

Revenons sur cette définition. Son principe général est de considérer l'écart minimal qui existe entre le support de m dans le contexte c, et son support dans les autres contextes. Ainsi, un motif m sera jugé pertinent s'il est fréquent dans c et peu fréquent dans tous les autres contextes. De plus, notons que cette définition ne considère que les contextes en conflit avec c pour déterminer la pertinence de m. En effet, il est inutile de tenir compte des contextes avec lesquels c n'est pas en conflit. Pour nous en convaincre, considérons le contexte $[\hat{a}g\acute{e}, 1]$ et un motif m. Il est inutile de mesurer si m est discriminant par rapport à $[jeune, 2]$ car jamais le classifieur n'aura à comparer ces deux contextes. Le seul contexte qui sera comparé à $[\hat{a}g\acute{e}, 1]$ est $[\hat{a}g\acute{e}, 2]$, i.e., le seul contexte en conflit avec $[\hat{a}g\acute{e}, 1]$.

Exemple 6. *Le tableau 1 montre, pour les motifs m_1, m_2 et m_3 spécifiques à $[\hat{a}g\acute{e}, 1]$, leur support dans les contextes $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$, ainsi que la mesure d'intérêt de chacun d'eux dans $[\hat{a}g\acute{e}, 1]$. Ainsi, si nous souhaitons sélectionner les deux meilleurs motifs pour décrire ce contexte, m_2 et m_1 (par ordre d'intérêt décroissant) seront sélectionnés.*

motifs	$[\hat{a}g\acute{e}, 1]$	$[\hat{a}g\acute{e}, 2]$	$disc_{[\hat{a}g\acute{e},1]}(m)$
m_1	1	0.8	0.2
m_2	0.9	0.65	0.25
m_3	0.9	0.8	0.1

TAB. 1 – *Sélection des meilleurs motifs pour le contexte $[\hat{a}g\acute{e}, 1]$*

3.3 Prise de décision

A ce stade, chaque contexte de notre hiérarchie est associé à un ensemble de k motifs séquentiels. Désormais, nous considérons une nouvelle puce à ADN dont la séquence correspondante S est associée à une valeur d^g sur la dimension guide D^g. Il s'agit donc de classer S dans un des contextes dits **candidats**, i.e., de la forme $[d^g, d^c]$, tel que d^c est une valeur sur la dimension de classement D^c, en exploitant les motifs séquentiels associés à chacun de ces contextes.

Score dans un contexte. Soit un contexte candidat c et la séquence S à classer. La séquence S obtient un score dans c défini comme suit :

$$score_c(S) = |\{m \in c | m \sqsubseteq S\}|.$$

En d'autres termes, le score de S dans un contexte c est le nombre de motifs associés à c qui sont inclus dans S.

Classement. Connaissant le score de S dans chaque contexte candidat, le processus de classification consiste à prendre une décision finale : *« Dans quel contexte candidat est classé S ? »*

Nous proposons ici de retenir le contexte candidat dans lequel S a obtenu le meilleur score [1]. La valeur d^c associée à S par le classifieur sur la dimension de classement D^c est telle que :

$$score_{[d^g, d^c]}(S) = \max_{x \in dom(D^c)} sup_{[d^g, x]}(S).$$

Exemple 7. *Considérons la séquence $S = \langle (G_4)(G_5G_6)(G_3)(G_1)(G_2) \rangle$ associée à une nouvelle puce à ADN dont nous voulons estimer le grade. Le patient est $\hat{a}g\acute{e}$. Par conséquent, nous recherchons les contextes candidats : $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$, i.e., qui respectent la valeur $\hat{a}g\acute{e}$ sur la dimension guide Age et qui possédent une valeur sur la dimension de classement $Grade$.*

Le tableau 2 présente pour les contextes $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$, les k meilleurs motifs de chaque contexte (pour $k = 3$).
Nous calculons le score de chaque contexte :

1. Dans le cas où le meilleur score est obtenu par plusieurs contextes candidats, alors le classifieur retourne les différentes valeurs possibles.

$[\hat{\mathbf{a}}\mathbf{g}\acute{\mathbf{e}}, \mathbf{1}]$	$[\hat{\mathbf{a}}\mathbf{g}\acute{\mathbf{e}}, \mathbf{2}]$
$\langle(G_1G_3)\rangle$	$\langle(G_4)(G_3)\rangle$
$\langle(G_5)(G_2)\rangle$	$\langle(G_5G_6)(G_2)\rangle$
$\langle(G_2)(G_4)(G_3)\rangle$	$\langle(G_3)(G_2)\rangle$

TAB. 2 – *Les motifs associés aux contextes* $[\hat{a}g\acute{e}, 1]$ *et* $[\hat{a}g\acute{e}, 2]$.

- $score_{[\hat{a}g\acute{e},1]}(S) = 1$
- $score_{[\hat{a}g\acute{e},2]}(S) = 3$

Par conséquent, le contexte retenu est $[\hat{a}g\acute{e}, 2]$ *car il s'agit du meilleur score, et le grade estimé de la puce à ADN est* 2*.*

4 Résultats expérimentaux

4.1 Description des données

Les données exploitées pour évaluer l'approche présentée proviennent de plusieurs enregistrements issus du NCIB [2] (National Center for Biotechnology Information). Un tri des enregistrements a été nécessaire afin de sélectionner un nombre suffisant de puces adéquates avec notre approche (i.e., qui contiennent toutes les informations contextuelles souhaitées). Le jeu de données utilisé pour les expérimentations est constitué de 649 puces à ADN, chacune étant associée à cinq dimensions contextuelles.

Gènes considérés. Les puces à ADN fournissent l'expression de milliers de gènes pour chacun des patients. Or, la plupart de ces gènes ne sont pas connus pour avoir une implication dans le cancer du sein. Nous nous sommes donc appuyés sur les 128 gènes identifiés par Sotiriou et al. (2006) pour leur implication dans cette maladie.

Dimensions contextuelles utilisées.

- **Grade** : le grade d'une tumeur constitue une des méthodes de diagnostic les plus anciennes et les plus utilisées. Elle permet d'établir trois grades de malignité grâce à l'évaluation de trois critères que sont la différenciation, le degré d'anisocaryose, ainsi que le nombre de mitoses. Valeurs : *grade 1, 2 ou 3*.
- **Age** : l'âge du patient lors du diagnostic. Valeurs : *50- (moins de 50 ans), 50-62 (entre 50 et 62 ans), 62+ (plus de 62 ans)*.
- **Taille** : la taille de la tumeur (en cm). Valeurs : *1.85- (moins de 1.85 cm), 1.85-2.45 (entre 1.85 et 2.45 cm), 2.45+ (plus de 2.45 cm)*
- **ER** : Récepteur des œstrogènes. Certaines tumeurs sont à récepteurs d'œstrogènes positifs (valeur *1*), c'est à dire que l'exposition des cellules cancéreuses à cette hormone va favoriser leur croissance. Au contraire, chez certaines personnes, le taux de croissance de la tumeur n'est pas affecté par cette exposition. Ces tumeurs sont à récepteurs d'œstrogènes négatifs (valeur *0*).

2. `http://www.ncbi.nlm.nih.gov`

– **Node** : invasion des ganglions lymphatiques. Cette dimension nous donne sur l'état des ganglions lymphatique, elle sera positive (valeur *1*) en cas d'invasion métastatique de ceux-ci, négative (valeur *0*) sinon.

4.2 Résultats

Les expérimentations effectuées visent à répondre à deux questions principales :

1. Quel est l'apport de l'utilisation d'une dimension guide dans les résultats de classification ?
2. En fonction des caractéristiques d'un nouveau patient, quel guide choisir pour maximiser les résultats ?

Dans un premier temps, nous construisons et évaluons les classifieurs obtenus à partir de chaque dimension guide, mais également lorsqu'aucune dimension guide n'est utilisée.

La qualité des classifieurs est évaluée par le biais du rappel et de la précision pour chaque contexte. La *précision* est le nombre de séquences correctement affectées à un contexte, divisé par le nombre total de séquences affectées à ce contexte. Le *rappel* est le nombre de séquences correctement affectées à un contexte, divisé par le nombre total de séquences appartenant réellement à ce contexte. Ces deux mesures sont utilisées pour calculer une *F-mesure* combinant rappel et précision. La F-mesure dans un contexte c est définie de la manière suivante :

$$F = 2 \cdot \frac{précision \cdot rappel}{précision + rappel}.$$

Guide		**Rappel**	**Précision**	**F-mesure**
sans guide		0.66	0.65	0.66
Age	50-	0.78	0.77	0.78
	50-62	0.80	0.78	0.79
	62+	0.81	0.80	0.80
Taille	1.85-	0,76	0,71	0.73
	1.85-2.45	0.81	0.80	0.81
	2.45+	0.76	0.73	0.74
ER	0	-	-	-
	1	0.71	0.70	0.70
Node	0	0.72	0.68	0.70
	1	0.89	0.91	0.90

TAB. 3 – *Tableau récapitulatif des performances des classifieurs construits (pour* $minSup = 0.8$ *et* $k = 5000$*)*

Le tableau 3 présente de manière résumée les performances obtenues pour chaque guide possible. Par exemple, le classifieur construit pour prédire le grade de cancer pour les personnes âgées de plus de 62 ans obtient une F-mesure de 0.8. Notons qu'aucun classifieur n'a été construit pour la valeur 0 sur la dimension ER. En effet, le contexte correspondant à cette valeur et au grade 1 ne contenant que 5 puces à ADN, il est impossible d'en extraire les motifs

séquentiels, et par conséquent de construire le classifieur associé. La première ligne montre les résultats obtenus sans utiliser de guide, i.e., sans exploiter les informations contextuelles. La F-mesure obtenue est alors 0.66. Nous constatons que les résultats sont toujours meilleurs lorsqu'un guide est utilisé, peu importe celui-ci. Ainsi, l'exploitation des informations contextuelles associées aux puces apporte une véritable amélioration des résultats.

De plus, ces résultats permettent de mettre en évidence les guides les plus avantageux pour la classification. Considérons l'exemple d'une puce associée aux informations contextuelles suivantes :

- Age : 53 ans,
- Taille de la tumeur : 2.5 cm,
- ER : 1,
- Node : 1.

Chacune de ces informations peut être utilisée comme guide. L'expert peut par conséquent choisir parmi les classifieurs correspondants celui qui sera le plus apte à classer correctement la puce. Dans ce cas, la dimension *Node* est celle qui lui offrira la meilleure F-mesure (0.90).

5 Conclusion

Dans cet article, nous avons présenté une approche de classification des puces à ADN reposant sur l'extraction de motifs séquentiels dans les puces à ADN ainsi que sur la prise en compte d'informations contextuelles liées aux patients. Nous avons montré l'intérêt de telles informations pour améliorer les performances de la classification afin de proposer des outils d'aide au diagnostic plus fiables. En exploitant ces connaissances, l'expert est en mesure de choisir le classifieur qui maximisera les chances de réussite pour le diagnostic.

Cependant, l'approche proposée ne permet de prendre en compte qu'une seule dimension contextuelle à la fois. Parmi les perspectives ouvertes par ces travaux, la première consiste à développer une méthode permettant de prendre en compte toutes les informations contextuelles afin de maximiser les performances en classification. De plus, le processus de choix d'un classifieur en fonction des informations contextuelles connues pourrait être automatisé (par exemple sous la forme d'un arbre de décision). Enfin, la méthode définie pourrait être exploitée pour d'autres maladies que le cancer du sein.

Références

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2).

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. S. P. Chen (Eds.), *Eleventh International Conference on Data Engineering*. IEEE Computer Society Press.

Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, et A. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America 96*(12), 6745.

Dougherty, E. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics 2*(1), 28–34.

Jaillet, S., A. Laurent, et M. Teisseire (2006). Sequential Patterns for Text Categorization. *International Journal of Intelligent Data Analysis (IDA) 10*(3).

Rabatel, J., S. Bringay, et P. Poncelet (2010). Contextual Sequential Pattern Mining. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*. IEEE Computer Society.

Rosenwald, A., G. Wright, W. Chan, J. Connors, E. Campo, R. Fisher, R. Gascoyne, H. Muller-Hermelink, E. Smeland, J. Giltnane, et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine 346*(25), 1937.

Salle, P., S. Bringay, et M. Teisseire (2009). Mining Discriminant Sequential Patterns for Aging Brain. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine : Artificial Intelligence in Medicine*, pp. 365–369. Springer-Verlag.

Simon, R. et K. Dobbin (2003). Experimental design of DNA microarray experiments. *Biotechniques 34*(Suppl 1), 16–21.

Sotiriou, C., S. Neo, L. McShane, E. Korn, P. Long, A. Jazaeri, P. Martiat, S. Fox, A. Harris, et E. Liu (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America 100*(18), 10393.

Sotiriou, C., P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, et al. (2006). Gene expression profiling in breast cancer : understanding the molecular basis of histologic grade to improve prognosis. *JNCI Cancer Spectrum 98*(4), 262.

Van De Vijver, M., Y. He, L. van't Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts, M. Marton, et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine 347*(25), 1999.

Zupan, B., J. Demsar, M. Kattan, J. Beck, et I. Bratko (2000). Machine learning for survival analysis : a case study on recurrence of prostate cancer. *Artificial intelligence in medicine 20*(1), 59–75.

Summary

Breast cancer is a major health problem and a challenge for biologists and health professionals. DNA microarrays now provide new tools to study the problems associated with this disease. In this paper, we propose to process data from DNA microarrays by discovering contextual sequential patterns (gene sequences ordered according to their expression level associated with a context). The goal is to provide aid in the diagnosis of a tumor grade. Our approach takes into account both the information contained in DNA microarrays (expressed through sequential patterns) but also of additional contextual information (e.g., patient age, tumor size, etc.) generally being associated with microarray data. The proposed approach has been evaluated on real data.