Identification de biomarqueurs et d'ARN non codants par des approches basées sur l'intelligence computationnelle

Anouar Boucheham

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Constantine 2 Abdelhamid Mehri
Faculté des Nouvelles Technologies de l'Information et la Communication
Département d'Informatique Fondamentale et ses Applications

Année : 2016
N° d'ordre : 068/2016

THÈSE

Pour l'obtention du diplôme de Docteur en 3éme cycle LMD
Option : Systèmes Complexes

Identification de biomarqueurs et d'ARN non codants par des approches basées sur l'intelligence computationnelle

---

Anouar BOUCHEHAM

Soutenue le 25 juin 2016 devant le jury composé de :

Pr. Nacereddine ZAROUR	Président	Université Constantine 2 - Abdelhamid Mehri
Pr. Mohamed BATOUCHE	Rapporteur	Université Constantine 2 - Abdelhamid Mehri
Dr. Sihem MOSTEFAI	Examinatrice	Université Constantine 2 - Abdelhamid Mehri
Pr. Abdelouahab MOUSSAOUI	Examinateur	Université de Sétif 1
Dr. Smaine MAZOUZI	Examinateur	Université de Skikda
Dr. Fariza TAHI	co-encadreur	Université d'Evry-Val-d'Essonne France

Remerciements

Je tiens d'abord à exprimer mes remerciements et gratitude au Seigneur Dieu de m'avoir donné la force et le courage de mener à terme ce projet de thèse.

J'aimerais adresser mes profonds respects et remerciements à Pr. Mohamed BATOUCHE, mon directeur de thèse, pour m'avoir accueilli, encadré, soutenu et prodigué de nombreux conseils tout au long de ces quatre années de thèse. Je lui suis particulièrement reconnaissant de m'avoir laissé une grande liberté scientifique, ce qui m'a permis de recevoir un apprentissage idéal et privilégié de la recherche. Je tiens également à exprimer ma gratitude au Pr. Souham MESHOUL pour son aide et l'attention dont elle m'a entourée tout au long de ce travail. Ainsi pour leurs remarques, leurs suggestions pertinentes, leurs encouragements et leurs confiance qui m'ont énormément servi et permis d'apprendre rapidement, je les remercie.

J'adresse mes sincères remerciements à Dr. Fariza TAHI, co-encadrante de cette thèse, qui m'a beaucoup appris et beaucoup apportée, sur le plan scientifique. Qu'elle soit consciente que ce travail ne serait pas là sans elle. Je la remercie pour sa disponibilité, son soutien et ses encouragements et de m'avoir offert l'occasion de passer une année de ma thèse au sein de son équipe au laboratoire IBISC. Mes remerciements vont aussi à tous les membres du laboratoire IBISC. Je remercie particulièrement les personnes qui m'ont permis de résoudre des problèmes aussi bien techniques qu'administratifs.

Je souhaite particulièrement adresser mes vifs remerciements aux membres du jury, Pr. Abdelouahab MOUSSAOUI, Dr. Smaine MAZOUZI , Pr. Nacereddine ZAROUR et Dr. Sihem MOSTEFAI pour avoir accepté de participer à mon jury de thèse.

Que le personnel de la Faculté NTIC trouve ici l'expression de ma sincère reconnaissance.

Mes plus vifs remerciements à mes parents et ma famille Boucheham. Je ne leur serai probablement jamais assez reconnaissant de m'avoir soutenu tout au long de mes études.

Je tiens également à remercier notre chère enseignante Nassira CHEKKAI pour ses remarques, ses conseils et son soutien tout au long de cette thèse.

Finalement, je ne saurai conclure sans remercier tous mes chers amis pour leur soutien inconditionnel, je cite particulièrement Hamada, Zaki, Kamel et Akram.

Résumé

Actuellement, le cancer prédomine comme le premier problème de santé dans le monde. La classification des cancers a toujours été fondée sur l'étude morphologique des tumeurs. Cependant, les tumeurs avec des apparences histologiques similaires peuvent présenter des réponses différentes au traitement, ce qui indique des différences de caractéristiques de la tumeur au niveau moléculaire. Ainsi, le développement d'une nouvelle méthode fiable et précise pour la classification des tumeurs est essentiel pour un diagnostic et un traitement plus efficace. Les biomarqueurs moléculaires fournissent de nouvelles façons permettant de comprendre le processus de la maladie et les moyens par lesquels les médicaments fonctionnent pour lutter contre la maladie. Au cours des dernières années, les chercheurs ont consacré un intérêt croissant à l'identification de biomarqueurs, en raison de son extrême importance en génomique et dans la médecine personnalisée.

Dans cette thèse, nous abordons le problème de la découverte de biomarqueurs à deux niveaux: génomique et transcriptomique. Nous nous intéressons d'abord au problème de la sélection des signatures moléculaires robustes et précises à partir des données d'expression génique qui s'appuie principalement sur les algorithmes de sélection de caractéristiques. L'objectif principal est d'atteindre de hautes performances de diagnostic assisté par ordinateur, en sélectionnant quelques gènes avec une forte puissance prédictive et une grande sensibilité aux variations dans les tests cliniques réels. À cette fin, nous étudions les méthodes basées ensemble et la coopération parallèle de métaheuristiques qui ont reçues une attention croissante en raison de leur pouvoir de donner une plus grande précision et stabilité qu'un algorithme unique peut atteindre. Dans cette direction, nous proposons une méthode parallèle de sélection de caractéristiques basée sur un méta-ensemble de filtres (MPME-FS) pour la découverte de biomarqueurs à partir des données d'expression génique. La deuxième méthode proposée pour la découverte de biomarqueurs est une méthode hybride wrapper / filtre de sélection de caractéristiques basée sur la coopération parallèle de métaheuristiques et un mécanisme à base de filtres pour l'initialisation et la réparation des solutions, appelé CPM-FS. Nous avons également proposé une méthode de sélection de gènes en deux étapes dont chacune est basée wrapper en utilisant la méthode précédemment proposée (CPM-FS) et une fonction de consensus qui prend en compte les dépendances entre gènes. Les expérimentations sur douze ensembles de données représentant différents types du cancer ont montré que

nos approches surpassent les méthodes récentes dans la littérature en terme de performance prédictive et fournissent également une sélection robuste à travers les différentes mesures de similarité. L'interprétation biologique des signatures sélectionnées indique que les méthodes proposées garantissent la sélection des gènes hautement informatifs pour le diagnostic du cancer.

Dans une deuxième partie de cette thèse, nous proposons une approche intégrative pour la prédiction des ARN non-codants (ARNnc) qui jouent un rôle important dans la régulation post-transcriptionnelle de gènes, soulignant leur importance en tant que biomarqueurs et leur impact sur le développement et la progression de nombreuses maladies. Dans l'approche proposée plusieurs types de propriétés génomiques et épigénomiques qui peuvent être utilisées pour caractériser ces molécules sont examinées. Nous développons un outil générique appelé IncRId qui permet de prendre en compte toutes les caractéristiques hétérogènes examinées de façon modulaire et facilement extensible et peut être utilisé et adapté pour prédire tout type d'ARNnc. Notre méthode permet également d'étudier la validité de chaque caractéristique dans chacune des espèces candidates. Par la suite, nous présentons un exemple d'application en se concentrant sur la prédiction d'ARNpi. Nous avons examiné et extrait un grand nombre de caractéristiques d'ARNpi de la littérature qui ont été observées expérimentalement chez plusieurs espèces. Nous avons implémenté ces caractéristiques dans un outil, appelé IpiRId, afin d'étudier la pertinence de chaque caractéristique dans chacune des espèces étudiées: humain, souris et mouche. Les résultats de prédiction d'IpiRId atteignent plus de 90% de précision, surpassant tous les outils existants. Le logiciel IpiRId et le serveur web de notre outil sont disponibles gratuitement pour les utilisateurs académiques à l'adresse: `https://tanuki.ibisc.univ-evry.fr/evryrna/IpiRId`

Mots clé: Apprentissage automatique; Bioinformatique; Médecine personnalisée; Analyse de l'expression génique; Identification de biomarqueurs; Classification des tumeurs; Sélection de caractéristiques; Apprentissage à noyau multiple; SVM; Sélection de gènes; Métaheuristiques parallèles inspirés de la nature; Méta-apprentissage; Données génomiques; Epigenomiques; Prédiction d'ARNnc, Prédiction d'ARNpi.

Abstract

Currently, cancer prevails as a prime health matter worldwide. Cancer classification has traditionally been based on the morphological study of tumors. However, tumors with similar histological appearances can exhibit different responses to therapy, indicating differences in tumor characteristics on the molecular level. Thus, the development of a novel, reliable and accurate method for the classification of tumors is essential for more successful diagnosis and treatment. Molecular biomarkers allow new ways of understanding disease processes and the manner in which medicines work to counteract disease. In the last few years, researchers have dedicated growing attention to biomarker identification given due to its extreme importance in genomics and personalized medicine.

In this thesis, we address the problem of biomarker discovery at two levels: genomics and transcriptomics. We are first interested in the problem of selecting robust and accurate signatures from gene expression data which relies heavily on the used feature selection algorithms. The main objective is to attempt high performance of computer-aided diagnosis (CAD), by selecting few genes with high predictive power and high sensibility to variations in real clinical tests. For that purpose, we have investigated ensemble-based methods and parallel cooperative metaheuristics which have received an increasing attention due to their power to give higher accuracy and stability than a single algorithm can achieve. Accordingly, we propose a parallel ensemble-based feature selection method based on meta-ensemble of filters (MPME-FS) for biomarker discovery from gene expression data. Then, we propose a hybrid wrapper/filter feature selection method based on the parallel cooperation of metaheuristics and a filter-based mechanism for both the initialization and the reparation of solutions, called CPM-FS. After that, we propose an ensemble-based wrapper gene selection method based on the previously proposed CPM-FS and a wrapper based consensus function in order to take into account genes dependencies. Experiments on 12 publicly available cancer datasets have shown that our approaches outperform recent state-of-the-art methods in term of the predictive performance. They also provide robust selection through the different similarity measures. Biological interpretation of the selected signature reveals that the proposed methods guarantee the selection of highly informative genes for cancer diagnosis.

In a second part of this thesis, we propose an integrative approach for the prediction of non-coding RNAs, which are molecules with an important

role in post-transcriptional gene regulation highlighting their importance as putative markers and their impact on the development and the progression of many diseases. In the proposed approach several types of genomic and epigenomic properties that can be used to characterize these molecules are examined. We have developed a generic tool called IncRId that allows taking into account all reviewed heterogeneous features in a modular and easily extensible way and could be used and adapted for predicting any type of ncRNA. Our method makes it possible to study the validity of each given feature in each of the candidate species. Then, we present an application example by focusing on the prediction of piRNAs. We reviewed and extracted a large number of piRNA features from the literature that have been observed experimentally in several species. We implemented these features in a tool, called IpiRId, to study the pertinence of each feature in each of the studied species: human, mouse and fly. IpiRId prediction results attain more than 90% accuracy, outperforming all existing tools. The IpiRId software and the web server of our tool are freely available to academic users at: `https://tanuki.ibisc.univ-evry.fr/evryrna/IpiRId`

Keywords: Machine learning; Bioinformatics; Personalized medicine; Gene expression analysis; Biomarker identification; Tumors classification; Feature selection; Multiple kernel learning; SVM classifier; Gene selection; Parallel nature-inspired metaheuristics; Meta-learning; Genomics data; Epigenomics; ncRNA prediction, piRNA prediction.

ملخص

في وقتنا الراهن، يتصدر السرطان لائحة المشاكل الصحية العالمية. تصنيف السرطان كان دائما مستندا على الدراسة المورفولوجية للأورام. غير أن الأورام ذات المظاهر النسيجية المتماثلة يمكن أن تُظهراستجابات مختلفة للعلاج، ما يشير إلى اختلافات في خصائص الورم على المستوى الجزيئي. لذلك، غدى إنشاء طريقة جديدة موثوقة ودقيقة لتصنيف الأورام أمرًا ضروريا من أجل تشخيصٍ وعلاج أكثر فعالية. المؤشرات البيولوجية الجزيئية تمنح طرقا جديدة لفهم تفاصيل المرض والطرق التي يواجه بها الطب هذا المرض.

إن ترجمة المؤشرات البيولوجية إلى اختبار تشخيصيٍّ ـ قائمٍ على مؤشراتٍ حيوية مهمة و مفيدة عياديًّا ـ تتطلب توليد وتحليل وتبادل نسبة معتبرة من البيانات والمعارف. في السنوات القليلة الماضية أولى البٍاحثون اهتماما متزايدا لتحديد المؤشرات البيولوجية تكنولوجيا نظرا لأهميتها القصوى في علم الجينوم والطب الشخصي. تطبيق المؤشرات الحيوية في مجال صحة الإنسان يمكّن من تحسين فهمنا للأمراض، باستطاعته كذلك تزويدنا بمعارف جديدة لآليات المرض. المؤشرات البيولوجية توفر أيضا وسيلة لتحسين الإدارة الصحية فيما يخص التشخيص المبكر للمرض وكذلك تقديم علاجات أكثر فعالية وأمنا.

في هذه الأطروحة، نعالج مشكل اكتشاف المؤشرات البيولوجية على مستويين: علم الجينوم و الإستنساخ. نهتم أولا بمشكلة اختيار التوقيعات القوية والدقيقة من بيانات التعبير الجيني الذي يعتمد بشكل كبير على خوارزمية اختيار الميزة المستخدمة. هدفنا الرئيسي هو بلوغ نسبة فعالية مرتفعة للتشخيصات بمساعد الحاسوب (CAD)، وذلك من خلال اختيارعدد قليل من الجينات مع قوة تنبؤية مرتفعة و حساسية عالية للمتغيرات في اختبارات عيادية حقيقية.

لهذا الغرض، قمنا باستخدام المناهج المعتمدة على الفرق والتعاون المتوازي للطرق التقريبية العامة التي تلقى اهتماما بالغا بسبب قدرتها على تقديم قدرٍ أكبر من الدقة والاستقرار مقارنة بالخوارزمية الواحدة. بالإضافة إلى أنها مؤهلة للتعامل مع العينات صغيرة الحجم وقاعدة البيانات المعقدة.

وفقا لذلك، نقترح نهجا مكررا لاختيار الميزات معتمدين على تجمع مُصفًّى (ME-FS) من أجل استكشاف المؤشرات البيولوجية عن طريق بيانات التعبير الجيني.نقترح في منهج ثانٍ ـ نُسميه MPME-FS - النموذج الموازي لمساهمتنا الأولى ME-FS ونوسعه ليشمل مرشحات وأنواع سرطان أخرى، هذا الحل الثاني المقترح لاكتشاف المؤشرات البيولوجية هو طريقة انتقاء ترتكز على تهجين المجمع/المرشح والذي يعتمد بدوره على التعاون المتوازي للطرق التقريبية العامة وآلية تصفية لتهيئة وتعويض كل من الحلين في كل طريقة تقريبية عامة، مسماة CPM-FS والتي ترمي إلى اختيار عدد محدد مسبقا من المؤشرات البيولوجية.

بعد هذا، نقترح طريقة اختيار الجينات على مرحلتين بحيث أن كلا منهما ترتكز على المجمع باستعمال المنهج المقترح سابقا (CPM-FS) ووظيفة توافقية تأخذ بعين الاعتبار التبعيات بين الجينات.

وقد أظهرت التجارب التي تم إجراؤها على إثنتي عشر مجموعة ـ تمثل أنواعا مختلفة من السرطان ـ أن مناهجنا تُنافِس و تتفوق على مناهج حديثة في الميدان، وذلك من حيث الفعاليّة التنبؤية. فضلا على هذا، تَهِبُ مناهجنا انتقاءً فعّالا بواسطة مقاييس التشابه المختلفة. الترجمة البيولوجية للتوقيعات المُختارة تُبين أن المناهج المُقترحة تضمن انتقاء الجينات الأغنى معلوماتيا من أجل تشخيص السرطان.

في الجزء الثاني من هذه الأطروحة، نقترح منهجا شاملا للتنبؤ بوحدات RNA غيرالمُشَفِرَة، وهي جزيئات تلعب دورا هاما في تنظيم مرحلة ما بعد نسخ الجينات مُؤكِدة أهميتها كمؤشرات بيولوجية وتأثيرها على تطور وتقدم العديد من الأمراض. في هذا المنهج المقترح تفحَّصنا عدة أنواع من الخصائص الجينومية و الاستنساخية التي يمكن استعمالها لوصف هذه الجينات. قمنا بتطوير أداة عامة تسمى lncRId والتي تسمح بأخذ كل الخصائص غير المتجانسة المفحوصة بعين الاعتبار وذلك بطريقة مُوحدة ومُوسَّعَة بسهولة ويمكن أن تُستخدم وتُكيَّف من أجل التنبؤ بأي نوع من RNAnc. طريقتنا تُمكّن من دراسة صحة كل ميزة في كل فصيلة مُرشّحة.

بعد هذا، نقدّم مثال تطبيقي مُركزين عل تنبؤ piRNA. اختبرنا واستخلصنا عددا كبيرا من صفات piRNA الموجودة في الميدان والتي كانت مُلاحَظةً تجريبيًّا عند فصائل كثيرة. نفذنا هذه الأخيرة في أداة تُسمى IpiRId بهدف دراسة ملاءمة كل صفة في كل من الفصائل المدروسة :البشر، الفئران والذباب. دقة نتائج تنبؤ IpiRId فاقت 90% متجاوزةَ كل الأدوات المتوفرة. برنامج IpiRId وخادم الويب الخاص بأداتنا متوفران مجانيا للمستخدمين الأكاديميين على الرابط التالي:

https://tanuki.ibisc.univ-evry.fr/evryrna/IpiRId

الكلمات المفتاحية:

التعلم الآلي، المعلوماتية الحيوية، الطب الشخصي، تحليل التعبير الجيني، تحديد المؤشرات البيولوجية، تصنيف الأورام، اختيار الميزات، التعلم المتعدد النواة، SVM، اختيار الجينات، الطرق التقريبية العامة المتوازية المستوحاة من الطبيعة، التعلم الفوقي، البيانات الجينومية، الإستنساخ، تنبؤ RNAnc ، تنبؤ piRNA.

Table des matières

Table des figures

Liste des tableaux

Chapitre

1 Introduction et problématique

Contenu du chapitre

La bioinformatique a progressé très rapidement depuis 20 ans, et va croître encore plus rapidement au cours des 20 prochaines années. Les progrès dans le domaine de la bioinformatique ont souvent apporté des évolutions importantes dans la médecine, en particulier, le traitement du cancer. Au cours des deux dernières décennies, le taux général de décès à cause de nouveaux cas de cancer reste aussi élevé, environ 49% dans l'ensemble, cette maladie génétique s'impose comme le premier problème de santé dans le monde. Ainsi, les efforts de la bioinformatique actuels se concentrent sur la découverte de biomarqueurs qui est l'élément clé de la médecine personnalisée, où la constitution génétique est utilisée pour guider les approches thérapeutiques. L'identification de biomarqueurs appropriés pour la détection précoce du cancer pourrait améliorer les soins de patients et a souvent entraîné des révolutions en médecine (Zhang *et al.*, 2011a).

La découverte de nouveaux biomarqueurs plus efficaces du cancer ou d'autres maladies génétiques est devenue une urgence, car leur développement et utilisation dans la pratique clinique va certainement conduire à des traitements sur mesure de ces maladies chez les patients (Nair *et al.*, 2014). La médecine individualisée ou personnalisée, aussi appelée thérapie ciblée, est un domaine, qui se concentre sur les différences entre les personnes et le potentiel de ce

dernier d'influer les résultats médicaux. Certes, il existe une différence entre les types du cancer, ils ne sont pas tous pareils. De sorte que le cancer d'une personne peut être catégorisé selon certains biomarqueurs qui sont présents ou absents, plus ou moins fréquents. De plus en apprenant davantage sur les cellules cancéreuses et leur milieu environnant, le nombre des sous-types de chaque cancer augmente. Les sous-types sont souvent basés sur les biomarqueurs qui permettent de distinguer le cancer basé sur certaines caractéristiques importantes, telles que l'agressivité de la maladie (biomarqueurs pronostiques) ou la réponse au traitement (biomarqueurs prédictifs) (Mäbert *et al.*, 2014). Par conséquent, on aura une plus grande probabilité de recevoir un traitement approprié et efficace pour un cancer particulier, contrairement à la méthode empirique utilisée dans le passé et actuellement pour déterminer le traitement.

Aujourd'hui, la découverte des biomarqueurs pour le dépistage et le diagnostic du cancer et d'autres maladies génétiques est devenue un point majeur dans les recherches sur le cancer. Les biomarqueurs sont également utilisés pour le pronostic, la surveillance de la progression du cancer, l'évaluation de l'efficacité du traitement et la prédiction de la récurrence de la tumeur. Parmi les différents types de biomarqueurs, on peut citer les biomarqueurs moléculaires (ADN, ARN, protéine ...) qui nous intéressent plus particulièrement car ils nécessitent beaucoup plus d'analyse, d'automatisation et de visualisation.

Dans le cadre de cette thèse, nous nous intéressons au problème de la découverte de biomarqueurs biologiques moléculaires (ADN et ARN) du cancer en utilisant des méthodes empruntées de l'intelligence computationnelle, afin de pouvoir proposer de nouvelles méthodologies bioinformatique pour leur identification. Le processus de la découverte nécessite l'exploitation et la prise en compte, dans un cadre d'apprentissage automatique supervisé, des caractéristiques biologiques qui peuvent discriminer ces molécules, ainsi que des données extraites de différentes nouvelles technologies de séquençages à haut débit (Puces à ADN, NGS, RNA-seq ...).

Principalement, nous traitons le problème de découverte de biomarqueurs à deux niveaux :

- Génomique (un biomarqueurs est un gène) à partir des données d'expression génique extraites des expériences des Puces à ADN.

- Transcriptomique (un biomarqueur est un ARN non-codant) à partir des données de séquence RNA-seq.

1.1 Découverte de biomarqueurs à partir de données d'expression génique

Plusieurs technologies avancées en génomique ont été développées ces dernières années (NGS, les puces à ADN et RNA-Seq ...), en particulier au cours de séquençage du génome humain, qui sont très utiles pour le diagnostic moléculaire, ce qui a mené à de nouvelles perspectives en biologie et ont conduit à la découverte de biomarqueurs (Zhang *et al.*, 2011a). Les puces à ADN ("Gene Expression Microarrays" (GEM)) sont la méthode la plus mature pour l'analyse génomique à haut débit. Elles sont utilisées pour surveiller et mesurer l'activité des gènes dans les tissus sains et malades à travers différentes populations. Elles peuvent aussi nous permettre d'identifier à la fois la protéine nécessaire à la fonction normale de la cellule et les anomalies produisant la maladie. En effet, ce qui est plus important est qu'ils arrivent à mesurer simultanément le niveau d'expression de tous les gènes de l'humain. Cela, en utilisant des milliers de sondes (ADN, ADNc) qui représentent des gènes spécifiques et sur la base du principe de l'hybridation d'acides nucléiques. Une fois une expérience de puce à ADN est terminée, l'analyse commence. Le plus grand défi dans les recherches en bioinformatique est la façon d'explorer et d'expliquer les données obtenues avec une grande précision.

La technologie des puces à ADN offre une plateforme idéale pour améliorer à la fois la découverte et la validation de biomarqueurs (Osl *et al.*, 2012). D'une manière générale, il ya plusieurs classes de problèmes étudiés dans GEM. Tout d'entre elles peuvent être divisées en trois classes; à savoir la classe prédiction qui utilise des approches d'apprentissage automatique supervisées, la classe découverte qui utilise des méthodes d'apprentissage automatique non supervisées (clustering) et, enfin, la classe comparaison des gènes qui utilise des méthodes d'apprentissage automatique en général (Golub *et al.*, 1999). Afin de proposer une approche plus compréhensible qui aide les médecins et les biologistes à détecter les principaux outils qui relient l'expression des gènes à des maladies. Nous nous concentrons dans ce travail sur la classe prédiction en utilisant des approches d'apprentissage automatique supervisées dont l'objectif est de sélectionner un sous-ensemble parcimonieux de biomarqueurs du cancer à partir de tout l'ensemble de données

d'expression de gènes, pour un type particulier du caner. Il n'y a pas de doute que l'utilisation de biomarqueurs appropriés pour les maladies génétiques a de nombreux avantages, à savoir la détermination de la prédisposition, la détection précoce, le diagnostic, l'évaluation du pronostic et de la réponse aux médicaments. Aujourd'hui, il ya un besoin urgent de biomarqueurs précis et robustes pour les cancers humains qui sont principalement des maladies génétiques causées par des mutations du génome (Lundblad, 2010).

Le fait que les données d'expression génique sont constituées d'un grand nombre de caractéristiques (gènes) et un nombre relativement petit d'échantillons, l'application directe de tout modèle d'apprentissage automatique sur des données de grande dimension est généralement inefficace. En outre, les ensembles de données d'expression de gènes contiennent un grand nombre de gènes redondants, bruyants et non pertinents. Ainsi, seul un sous ensemble de gènes est biologiquement pertinent et donne ainsi une grande précision pour le diagnostic du cancer. De plus, la présence de nombreuses caractéristiques affecte non seulement les performances de prédiction, mais également le temps de calcul des algorithmes de classification tel que K-plus proches voisins, machines à vecteurs de support (SVM), entre autres (Somorjai *et al.*, 2003). Pour éviter le problème de la malédiction de la dimensionnalité, il devient alors nécessaire de sélectionner un petit sous-ensemble de caractéristiques / gènes qui peuvent séparer les patients sains des patients atteints du cancer ou de façon plus générale, des gènes qui sont pertinents, non-redondants et discriminants pour une maladie génétique particulière. Ces gènes sont appelés biomarqueurs, gènes informatifs ou gènes différentiellement exprimés. Par conséquent, nous avons besoin des techniques de réduction de dimensionnalité, qui identifient un petit ensemble de gènes représentants l'information la plus discriminante de l'ensemble original de gènes afin d'obtenir une meilleure qualité d'apprentissage. Cette étape joue un rôle central dans le domaine de l'apprentissage automatique et plus particulièrement dans la tâche de classification.

Typiquement, l'identification de biomarqueurs et la classification du cancer sont deux problèmes étroitement liés. Du point de vue d'apprentissage automatique, la découverte de biomarqueurs est un problème de sélection de caractéristiques ("feature selection") et le diagnostic du cancer est un problème de classification supervisée où chaque classe est le phénotype d'un cancer spécifique. La sélection de caractéristiques est un axe en plein croissance dans l'apprentissage automatique. Elle est définie comme le processus d'identification et d'élimination des gènes inutiles et redondants à partir des

données d'entrainement. Ainsi, elle vise à sélectionner un sous-ensemble de caractéristiques (gènes) qui contribuent le plus dans la classification de cancers, en éliminant les gènes qui ne sont pas discriminants (Bolón-Canedo *et al.*, 2014b). En outre, la classification ou la prédiction du cancer fait référence à la procédure de la construction d'un modèle d'apprentissage (classificateur) en utilisant en entrée les gènes significatifs déjà identifiés dans l'étape de sélection. Cela est suivi par l'étape de prédiction qui utilise le model entrainé afin de pouvoir affecter de nouvel échantillon aux sous-types appropriés d'une maladie (Wu *et al.*, 2012).

1.2 Identification et classification des ARN non-codants. Application sur les ARNs Piwi

L'identification des interactions survenant au niveau moléculaire joue un rôle crucial pour la compréhension du vivant. Dans ces dernières années, les études de séquençage à haut débit ont montré que le génome humain est largement très transcrit (plus de 74,7% du génome humain) en milliers d'ARN qui ne codent pas pour une protéine (RNAnc). Tandis que pas plus de 2% de l'ensemble du génome code pour une protéine et joue le rôle de structure transitoire de l'information génétique par l'ARN messager (RNAm) (Djebali *et al.*, 2012).

Traditionnellement, les ARN non-codants sont divisés en fonction de leur taille, en petits ARNnc qui sont plus courts que 200 nucléotides (nt), et longues ARNnc d'au moins 200 nt de longueur, sur la base de protocoles de purification d'ARN (Kapranov *et al.*, 2007). Les petit ARNnc ont été largement étudiés au cours de la dernière décennie, et ont été montré à jouer un rôle dans la régulation des gènes au niveau transcriptionnel et post-transcriptionnel, en majorité, par appariement de bases spécifiques avec leur ARN / ADN cible (Huang *et al.*, 2013). Les longues ARNnc sont moins caractérisés, avec une structure et fonction exactes connues seulement pour plusieurs d'entre eux (Fatica and Bozzoni, 2014). Cependant, un nombre croissant d'études ont révélé qu'ils peuvent être impliqués dans la régulation de toutes les étapes de l'expression génique, soulignant l'importance d'une connaissance plus approfondie sur les longues ARNnc et leur pertinence fonctionnelle. Donc, Les ARN non-codants ont plusieurs rôles dans la cellule à savoir, la régulation du génome et la régulation post-transcriptionnelle, soulignant ainsi leur importance putative comme marqueurs et leur impact au cours du développement et la progression de nombreuses maladies, en par-

ticulier le cancer. Pour résumer, nos connaissances sur les familles d'ARNnc et leurs rôles est toujours en expansion. Les ARN non-codants sont l'un des points clés pour la compréhension du métabolisme des cellules et des organismes. Ils présentent un intérêt croissant par leur importance dans le fonctionnement des cellules, mais également par le développement des technologies de séquençage, qui permettent de les détecter bien plus rapidement qu'auparavant avec un moindre coût.

Les méthodes expérimentales d'identification des ARNnc sont très couteuses en temps comme en argent et assez difficiles à mettre en place. Donc, les approches bioinformatiques constituent ici un moyen intéressant pour orienter cette recherche. De manière générale, en bioinformatique la découverte de biomarqueurs au niveau transcriptomique se fait directement sur des milliers de petites et longues séquences d'ARN obtenus par la technologie RNA-seq, afin de les identifier et les classifier (codant ou non codant, long ou petit, miRNA, snoRNA ou piRNA ...). Cette identification sert à reconnaitre le type d'ARN et par conséquent son rôle en tant que biomarqueur dans la cellule.

Dans le cadre de cette thèse nous nous intéressons à la prédiction et l'identification des ARN non-codants (qui ne codent pas pour une protéine) à partir des données RNA-seq. Le processus d'identification par apprentissage se base principalement sur les caractéristiques qui peuvent discriminer chacune de ces molécules. L'objectif principal est d'arriver à proposer une nouvelle méthodologie bioinformatique basée sur l'apprentissage automatique supervisé qui permet l'identification précise des ARN non-codants, en termes de sensibilité et spécificité. Dans un second temps, appliquer la méthodologie proposée pour la prédiction des ARNs piwi (ARNpi), une classe d'ARN non-codants récemment découverte, qui interagissent avec les protéines PIWI (appartenant à la famille des Argonautes) pour inhiber les éléments transposables et ainsi protéger l'intégrité du génome. Les ARNs Piwi quant à eux ne présentent aucune structure particulière et leur mode d'expression peut varier d'une espèce à l'autre. De plus, aucun motif consensus n'existe, ce qui les rend difficiles à prédire.

1.3 Liste des contributions

Dans cette thèse nous proposons plusieurs nouvelles methodes et approches informatique et bioinformatique pour la prédiction et l'identification de biomar-

qeurs dans deux niveaux: génomique et transcriptomique qui se résument dans:

- La proposition d'une nouvelle approche parallèle de sélection de caractéristique basée sur la notion du meta-learning et un ensemble de filtres (Information-Gain, ReliefF ...) pour la découverte de biomarqueurs à partir des données d'expression génique.
- La proposition d'une approche hybride (wrapper /filter) basée sur la coopération parallèle des meta-heuristiques et un mécanisme de réparation à base de filtre, pour la découverte de biomarqueurs à partir des données d'expression génique.
- La proposition d'une méthode basée ensemble qui effectue la sélection en deux étapes dont chacune est basée wrapper.
- La proposition d'une nouvelle méthodologie bioinformatique intégrative et générique pour la prédiction de tous les types des ARNs non codant ainsi qu'un outil modulaire et facilement adaptable et extensible appelé "IncRId".
- La réalisation d'une étude approfondie sur les ARN piwi, et la proposition d'un outil, appelé "IpiRId", basé sur la méthodologie déjà proposée et spécifique pour l'identification des ARN piwi à partir de plusieurs types de données génomique et épigénomique.

1.4 Organisation de la thèse

Cette thèse est composée de six chapitres et deux parties dont nous présentons une brève description ci-dessous:
La première partie qui est composée de deux chapitres est consacrée à l'état de l'art:
Le **Chapitre 2** débute par l'introduction des différentes notions fondamentales biologiques et bioinformatiques. Une seconde partie est consacrée aux notions de base liées à la médecine personnalisée y compris les biomarqueurs et le rôle qu'ils jouent dans un processus de traitement personnalisé du cancer.
Le **Chapitre 3** présente un état de l'art sur des techniques informatique nécessaire à la bonne compréhension des méthodes proposées dans cette thèse. D'abord, nous introduisons le principe et les paradigmes de bases de l'intelligence computationnelle. Après cela, les éléments clés de l'apprentissage automatique supervisé sont définis, ainsi qu'une description des algorithmes

de classification usuels. Ensuite, nous introduisons de façon informelle le problème de sélection de caractéristique ainsi que ses concepts de base. En fin, nous présentons un état de l'art des méthodes de sélection de caractéristiques en mettant l'accent sur les approches proposées pour la découverte de biomarqueurs et montrant les avantages et les inconvénients de chaque catégorie de méthodes. La deuxième partie introduit les différentes contributions et méthodes proposées pour la découverte de biomarqueurs au niveaux génomique et transcriptomique.
Le **Chapitre 4** présente d'abord les différents aspects qu'il faut prendre en compte lors de la proposition d'une méthode basée ensemble pour la sélection de gènes. Ensuite, nous présentons une nouvelle approche parallèle de sélection de caractéristique basée sur la notion du meta-learning et un ensemble de filtres pour la découverte de biomarqueurs à partir des données d'expression génique.
Le **Chapitre 5** est consacré à la présentation de deux méthodes proposées pour la sélection de gènes les plus représentatifs pour le diagnostic du cancer. La première, appelée CPM-FS, est une méthode hybride wrapper/filter basée sur la coopération parallèle des méta-heuristiques et un mécanisme de réparation et d'initialisation à base de filtres. La deuxième, appelée ECPM-FS, sélectionne les gènes à deux étapes dont chacune est basée wrapper.
Le **Chapitre 6** est consacré à la présentation de la méthodologie bioinformatique intégrative et générique pour la prédiction de tous les types des ARNs non codant, ainsi qu'à la description de l'outil correspondant « IncRId ». Dans une seconde partie du chapitre, nous présentons une application exemple de la méthodologie proposée sur les ARN piwi, et nous décrivons l'outil « IpiRId » spécifique pour l'identification des ARN piwi à partir de plusieurs types de données génomique et épigénomique. La thèse se termine par une synthèse de nos différentes contributions et publications, et quelques perspectives de recherche.

Partie I

Etat de l'art

Nous présentons dans cette première partie les différentes notions biologiques et informatiques nécessaires à la compréhension des contributions de cette thèse. Il n'est certainement pas possible de couvrir tous les détails sur les sujets concernés afin de fournir une thèse autonome; nous présentons donc que les concepts que nous trouvons crucials pour les deux aspects informatique et biologique. Cette partie est divisée en deux chapitres: le premier est consacré aux concepts biologiques sur lesquels s'appuient nos travaux, et le second couvre les concepts de l'intelligence computationnelle.

Chapitre

2 Bioinformatique et médecine personalisée

Contenu du chapitre

2.1 Biologie moléculaire

Tout être vivant est composé de cellules comme une maison est faite de briques. La cellule constitue l'unité de base de tout organisme, des bactéries aux êtres humains en passant par les plantes et les champignons. Chaque

être humain est composé de quelque 100′000 milliards de cellules. En effet, la biologie moléculaire est apparue au XXe siècle, à la suite de l'élaboration des lois de la génétique, de la découverte des chromosomes et de l'identification de l'ADN comme support de l'information génétique, au croisement de la génétique, de la biochimie et de la physique.

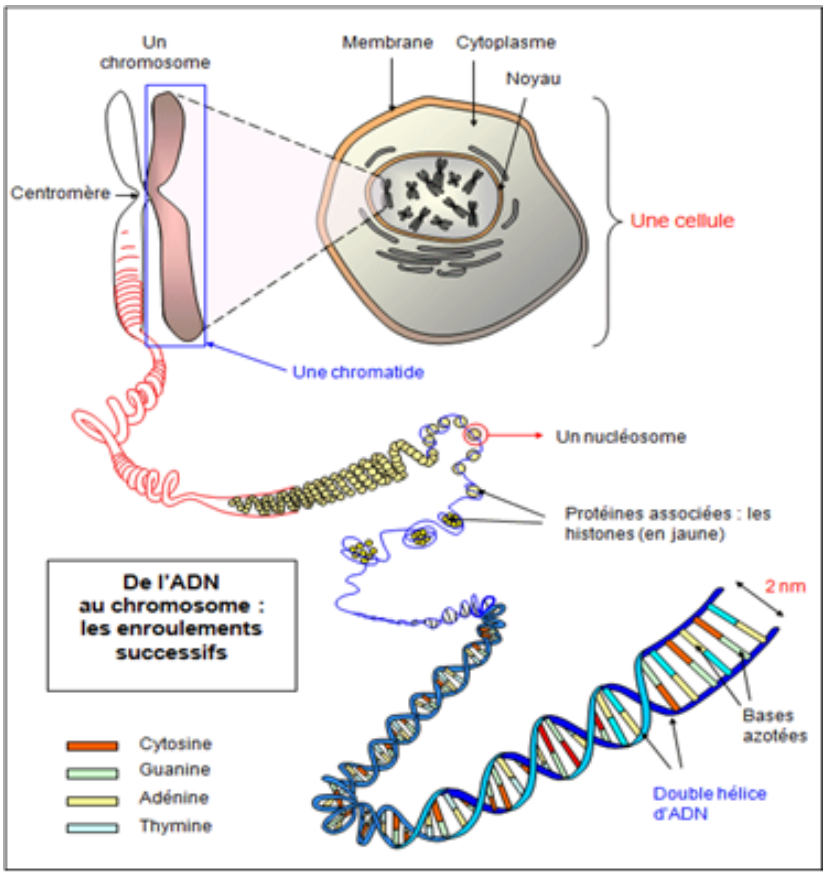


Figure 2.1: Les élements qui composent une cellule.

La biologie moléculaire est une branche de la science concernant l'activité biologique au niveau moléculaire. Le domaine de la biologie moléculaire se chevauche avec la biologie et la chimie et en particulier, de la génétique et de la biochimie. La biologie moléculaire examine les mécanismes moléculaires, à l'origine des processus, tels que la réplication, la transcription, la traduction et la fonction des cellules. Une façon simple de décrire la biologie moléculaire est qu'elle sert à comprendre comment les gènes sont transcrits en ARN et comment l'ARN est ensuite traduit en protéine. Cependant, cette image simplifiée est actuellement à être reconsidérer et réviser en raison de

nouvelles découvertes concernant les rôles de l'ARN (Lodish *et al.*, 2005). Le terme « biologie moléculaire » désigne également l'ensemble des techniques de manipulations d'acides nucléiques (ADN, ARN).

Dans ce qui suit nous résumons les principaux composants moléculaires de la cellule (ADN, ARN, protéine, gène...) qui constituent les fondations chimiques de la structure et des fonctions cellulaire (voir Figure 2.1).

2.1.1 Acide aminé (AA)

Une molécule organique possédant à la fois l'acide carboxylique ($-COOH$) et un aminé basique ($-NH_2$) attachés au même atome de carbone, comme montré dans la Figure 2.2. Les acides aminés sont les principaux éléments constitutifs des protéines et des enzymes. Ils sont incorporés dans les protéines par l'ARN de transfert en fonction du code génétique tandis que l'ARN messager est décodé par les ribosomes. La teneur en acide aminé détermine les propriétés spatiales et biochimiques de la protéine ou enzyme pendant et après la construction final d'une protéine. Plus de 50 AA ont été découverts, 20 sont essentiels pour la fabrication des protéines qui sont de longues chaînes d'acides aminés liés (Rittner and McCabe, 2009).

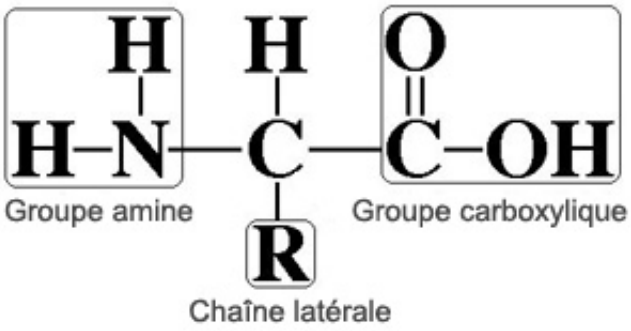


Figure 2.2: Acide aminé (AA).

2.1.2 Acide désoxyribonucléique (AND)

Un polymère linéaire de masse moléculaire élevée, composée de nucléotides contenant 2-désoxyribose et reliée entre les positions $3'$ et $5'$ par des groupes

phosphodiester[1]. L'ADN est une molécule présente dans toutes les cellules qui peut être en simple brin ou double brin, et qui contient l'information génétique transmise entre générations. Un brin simple (aussi appelé polynucléotide) est un polymère linéaire composé de 4 nucléotides: adénosine(A), cytosine (C) guanine (G) et thymine (T). On représente un polynucléotide par une séquence orientée de lettres:

$$5' - A - T - T - C - A - G - G - C - A - T - T - A - G - C - 3'$$

L'ADN a eu beaucoup d'attention depuis la découverte de sa structure de double-hélice ou forme tordue par James Watson et Francis Crick en 1953. Cette découverte a révélé que l'ADN porte l'information génétique pour le développement et le fonctionnement des organismes vivants et qui se retrouve sous forme stable de double brin, comme montré dans la Figure 2.3.

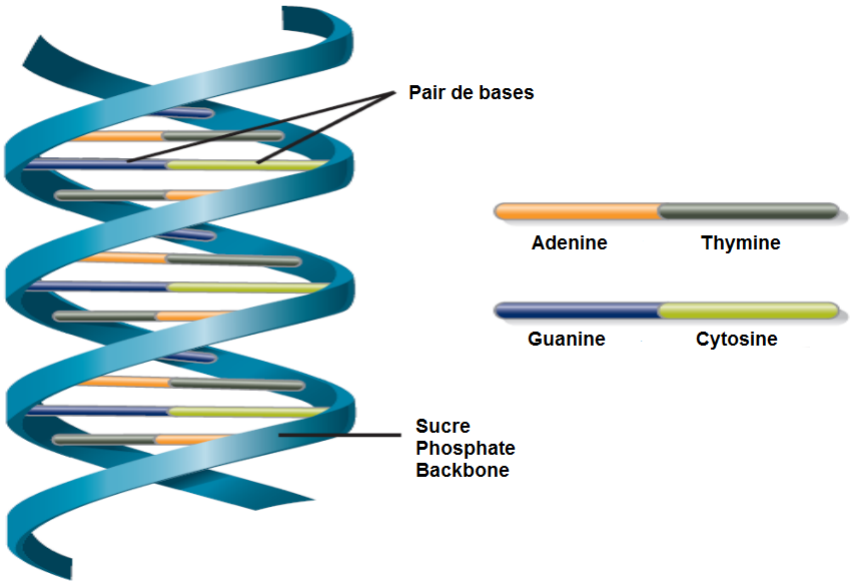


Figure 2.3: Acide désoxyribonucléique (AND).

L'ADN contient l'information qui demande aux cellules de développer des caractéristiques spécifiques, qui leur permettent d'exécuter des fonctions exactes dans le corps. Par exemple, les cellules nerveuses sont conçues pour

[1]Une liaison phosphodiester lie un groupe phosphate et les deux carbones 3 et 5 de deux molécules de sucre par deux liaisons ester. La liaison « esther » s'établit entre un groupe alcool (OH) et un groupe carboxyle ($COOH$) avec élimination d'une molécule d'eau.

communiquer l'information, et les cellules cancéreuses sont conçues pour se développer et replier. En outre, l'ADN porte les gènes qui composent l'information héréditaire qui est passée de génération en génération. Aucunes deux personnes n'ont exactement le même ADN, à l'exception des jumeaux identiques. Un brin ou un ordre de l'ADN chez l'humain peut se composer de jusqu'à 2 millions de bases azoté (A, C, G et T), ces ordres s'appellent les gènes. Les gènes peuvent contenir des milliers de bases nucléotidiques. Dans le noyau de la cellule, l'ADN est étroitement trouvé avec des protéines en structures appelées chromosomes.

2.1.3 L'acide ribonucléique (ARN)

L'acide ribonucléique (ARN) est une macromolécule formée par la polymérisation de nombreux nucléotides, tout comme l'ADN. Les principales différences sont que le ribose remplace le désoxyribose dans l'ARN, qui n'est pas double brin comme l'ADN, mais plutôt un simple brin. Les nucléotides qui le composent sont l'adénine (A), la guanine (G), la cytosine (C) et l'uracile (U), ce dernier est la contrepartie de (T) la thymine. Les trois types les plus importants de l'ARN dans la cellule sont des ARN messager (ARNm), ARN de transfert (ARNt) et l'ARN ribosomique (ARNr). La structure d'une molécule d'ARN est également déterminée par sa séquence d'ADN dérivée. Si les protéines sont le matériel, l'ARN est le logiciel qui controle la façon dont les gènes sont exprimés pour fabriquer des protéines. L'ARN est unique dans sa capacité à stocker et à transmettre des informations ainsi que des processus de cette information.

ARN codant (ARNm)

Classiquement les ARN peuvent être classés en ARN messagers (ARNm), qui sont traduits en protéines et ARN non-codants (ncRNAs) qui ne codent pas pour une protéine. L'ARNm est l'intermédiaire de courte durée dans le transfert de l'information génétique de l'ADN aux protéines. L'ARNm est transporté hors du noyau et est traduit en protéine sur les ribosomes cytoplasmiques. Le transcriptome est l'ensemble des molécules d'ARNm d'une cellule, un tissu ou un organisme. La transcription conserve tout le contenu de l'information de la séquence d'ADN qu'il a été transcrit à partir, puisque l'ARN a les mêmes caractéristiques d'appariement de bases.

ARN non codant (ARNnc)

Les ARNnc produisent des molécules d'ARN fonctionnels plutôt que codants pour des protéines et comprennent les ARN de transfert (ARNt) et des ARN ribosomique (ARNr). Les ARNr sont des molécules trouvées dans tous les organismes vivants fortement structurés et conservés et sont bien établis comme marqueurs phylogénétiques. Au cours des deux dernières décennies, plusieurs ARNnc ont émergé, ayant un large éventail de fonctions, de la structure par le biais réglementaire catalytique. Une catégorie dominante est celle des petits ARNnc à savoir les microARN (miARN), les snoARN, les ARNpi ...ect.

2.1.4 Protéine

Une protéine est une macromolécule composée d'une ou plusieurs chaînes d'acides aminés (les éléments de base) liés entre eux par des liaisons peptidiques. Nos corps s'alimentent des protéines à partir des aliments que nous mangeons en acides aminés individuels. Ces acides aminés sont ensuite réassemblés en protéines spécifiques que notre corps a besoin, y compris la structure et la fonction cellulaire, ainsi que la réglementation des tissus et organes du corps. Ils représentent l'une des plus importantes classes de molécules dans les organismes vivants. Leurs fonctions incluent la catalyse de processus métaboliques sous la forme d'enzymes, ils jouent aussi un rôle important dans la transmission du signal, les mécanismes de la défense et le transport de molécules.

2.1.5 Gène

Un gène est une séquence d'ADN chromosomique qui est nécessaire pour la production d'un produit fonctionnel: un polypeptide ou une molécule d'ARN fonctionnel. La taille des gènes varie de la petite ($1,5kb$ pour globine) à la grande taille (environ $2.000kb$ pour "Duchenne gene" de la dystrophie musculaire). Un gène comprend non seulement les séquences de codages réels, mais également des séquences de nucléotides adjacentes nécessaires à l'expression appropriée des gènes pour la production d'une molécule d'ARNm normal. L'ARNm mature est d'environ 1/10 de la taille du gène d'où il est transcrit. Le même brin d'ADN d'un gène est toujours traduit en ARNm de sorte qu'un seul type d'ARNm est effectué pour chaque gène. Les gènes sont souvent décrits comme des plans de la vie et de transmettre des traits hérités (caractéristiques) d'une génération à l'autre (Jain, 2015).

Structurellement, un gène représente l'unité de base du matériel héréditaire qui est une séquence ordonnée de bases de nucléotides et qui code pour une chaîne polypeptidique (par l'intermédiaire de l'ARNm). Il existe dans tout organisme vivant et porté par une ou plusieurs molécules d'ADN présentes dans chaque cellule. C'est l'ordre dans lequel ces nucléotides sont enchaînés et qui permet de coder l'information génétique de chaque individu. Les gènes sont au centre de la génomique, une discipline qui étudie la structure, le fonctionnement et l'évolution des génomes.

2.1.6 Chromosome

Un chromosome se compose de l'ADN étroitement emballé et supporté par des protéines appelées histones. Chaque chromosome humain est une longue molécule linéaire d'ADN double brin (sauf le chromosome mitochondrial) dont la taille varie de 50 à 250 millions de paires de bases, situé dans le noyau de la cellule. Un chromosome moyen contient entre 2000 et 5000 gènes à l'intérieur de 130 millions de paires de bases et est égale à environ $130cm$ de matériau génétique. Il ya environ 400 millions de nucléotides dans un chromosome humain, mais seulement environ 2% d'entre eux codent pour des protéines, le reste peut jouer différents rôles tels que la régulation de l'expression des gènes.

En effet, les différents organismes ont différents nombres de chromosomes. Par exemple, les cellules de l'humain ont 23 paires de chromosomes: 22 paires de chromosomes numérotées, appelés les autosomes et qui sont les mêmes dans les mâles et les femelles, ainsi qu'une paire de chromosomes sexuel, qui permet de faire la différence entre les mâles (un chromosome X et un autre Y) et les femelles (deux chromosomes de X).

2.1.7 Génome

C'est l'assemblage complet des chromosomes et des gènes extra-chromosomiques d'une cellule et qui représente la portion complète de l'ADN d'un organisme ou l'ensemble complet de gènes partagés par les membres d'un organe de reproduction tels qu'une population ou espèce. Le génome est transmis de génération en génération et se réfère également au support physique de cette information génétique, c'est-à-dire la macromolécule d'ADN.

2.1.8 Le code génétique

La séquence de bases nucléotidiques du "code génétique" dans un gène particulier reflète une séquence spécifique d'acides aminés dans le polypeptide produit par le mécanisme de synthèse des protéines. La colinéarité entre la molécule d'ADN et la séquence de protéine est obtenue à l'aide du code génétique. En tout, il existe quatre positions possibles (A, T, C et G). Ainsi, pour les trois bases, il ya (4^3) ou 64 combinaisons possibles de triplets et qui constituent le code génétique.

2.1.9 Promoteur (séquence promotrice)

Une séquence promotrice est une région située à proximité d'un gène et indispensable à la transcription, sur laquelle se fixe l'ARN polymérase. Les séquences promotrices sont situées en amont du site d'initiation de la transcription.

2.1.10 L'information génétique : du gene à la protéine

La synthèse des protéines est l'acte par lequel une cellule assemble une chaîne protéique en combinant des acides aminés isolés présents dans son cytoplasme, guidé par l'information contenue dans l'ADN. Elle se déroule en trois étapes: la transcription de l'ADN en pré-ARN, l'épissage du pré-ARN à l'ARN messager et la traduction de l'ARN messager en une protéine.

La transcription

Les gènes sont transcrits en pre-ARN par un ensemble complexe de molécules (ARN polymérase). Durant la transcription, la lettre T (thymine) est remplacée par la lettre U (Pour Uracil) (Gonzalez de Castro *et al.*, 2013).

Epissage

Le Pre-ARN est représenté par des alternations de segments de séquences appelés exons et introns. L'épissage est la concaténation des exons et suppression des introns pour former l'ARNm (ou simplement l'ARN). Les Exons sont les parties du pre-ARN qui seront traduites en protéines (Gonzalez de Castro *et al.*, 2013).

Traduction

Au cours de la troisième étape de l'expression des gènes, connue sous le nom de traduction, l'information qui est contenue dans l'ARNm est traduite dans une autre langue en une structure à l'intérieur du cytoplasme appelée un ribosome. Le ribosome lit la séquence de bases nucléotidiques, avec trois nucléotides codant pour un acide aminé particulier, suivant le code génétique Xiong (2006). Cette séquence de trois nucléotides est appelée un codon. Les acides aminés sont les blocs de construction des protéines. Un type d'ARN appelé ARN de transfert (ARNt) assemble ensuite les acides aminés dans l'ordre de leur lecture par le ribosome. Les protéines sont simplement de longues chaînes d'acides aminés qui se replient sur différents modes d'enroulement ou en fonction de leur longueur et de la séquence d'acides aminés.

2.1.11 Technologie des puces à AND

Les puces à ADN, également appelées "DNA arrays", DNA chips, DNA microarrays ou microarrays, ont été développées au début des années 1990. Elles permettent de mesurer simultanément et quantitativement l'expression de plusieurs milliers de gènes à partir d'un support solide de taille réduite (quelques cm^2). Depuis leur apparition, les puces à ADN sont devenues des outils majeurs pour la recherche en biologie fondamentale et clinique (Quackenbush, 2006). Un ensemble de données d'expression de gène rassemble l'ensemble des valeurs d'expression d'une série d'expériences de puces à ADN, avec chaque ligne qui représente un échantillon et plusieurs milliers de gènes représentés en colonnes (Hastie *et al.*, 2005).

La technologie des puces à ADN, connaît à l'heure actuelle un essor important et suscite un formidable intérêt dans la communauté scientifique. En plein essor depuis le début des années 2000, cette technologie est devenue très accessible via sa commercialisation par des sociétés comme Affymetrix. Grâce à cette méthodologie, la mesure simultanée du niveau d'expression de plusieurs milliers de gènes dans des dizaines de conditions différentes, physiologiques ou pathologiques, est aujourd'hui techniquement possible. Le potentiel de cette méthodologie est énorme, mais la masse des résultats qu'on peut avoir à partir de ces expériences est considérable (plusieurs dizaines de milliers de résultats peuvent être obtenus simultanément) et leur exploitation par le biais de programmes informatiques n'est encore qu'à ses débuts.

2.1.12 L'expression de gènes

L'expression de gènes est le processus par lequel un gène se fait activé dans une cellule pour fabriquer des ARN (acide ribonucléique) et des protéines. L'activité d'un gène, que l'on appelle l'expression de gènes signifie que l'ADN est utilisé comme modèle pour produire une protéine spécifique. Seul un petit nombre de ces gènes, environ 15.000, est exprimé dans une cellule humaine typique, mais les gènes exprimés varient d'une cellule à l'autre. L'expression génique peut être détectée par différentes techniques à savoir les puces à ADN (voir Figure 2.4).

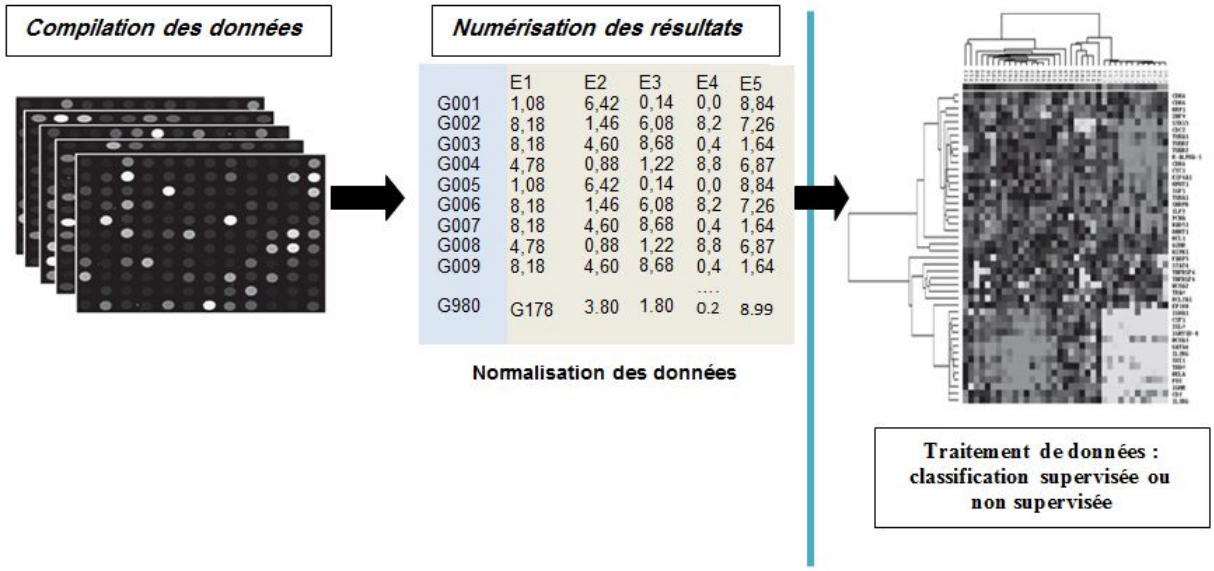


Figure 2.4: Représentation des données des puces à ADN

Toutes les fonctions de cellules, tissus ou organes sont commandés par l'expression de gènes différentielle. La connaissance des gènes qui sont exprimés dans les tissus sains et malades nous permettrait d'identifier à la fois la protéine nécessaire à la fonction normale et les anomalies qui causent la maladie. Cette information aide dans le développement de nouveaux tests de diagnostic pour diverses maladies ainsi que de nouveaux médicaments pour modifier l'activité des gènes ou des protéines affectées. Un Mauvais fonctionnement de gènes est impliqué dans la plupart des maladies, pas seulement ceux hérités.

Ces études ouvrent de nouvelles perspectives pour identifier les gènes et les biomarqueurs complexes pour le diagnostic des maladies et l'évaluation

d'efficacité et de toxicité des médicaments. Il existe trois types de problèmes principalement étudiés à partir des ensembles d'expression génique (Zhang *et al.*, 2011a):

- **Classe prédiction:** la détermination de l'état fonctionnel de la cellule en fonction de niveau d'expression de gènes.
- **Sélection de gènes:** l'identification des gènes en corrélations avec l'état fonctionnel de la cellule suspect.
- **Classe découverte:** l'analyse des groupes (clusters) de gènes co-exprimés et fonctionnellement corrélés.

2.2 Qu'est-ce que la bioinformatique ?

Avant l'apparition de la bioinformatique, seulement deux façons pour réaliser des expériences biologiques étaient disponibles: au sein d'un organisme vivant (dite "in vivo") ou dans un environnement artificiel (dite "in vitro", ou "in glass"). Prenant plus loin l'analogie, nous pouvons dire que la bioinformatique est en effet la biologie "in silico", des puces de silicium sur lesquelles les microprocesseurs sont construits (Claverie and Notredame, 2011).

Aujourd'hui, la quantité des données biologiques accumulées dans les laboratoires explose. De ce fait, elles ne peuvent plus être analysées « à la main » comme autrefois, donc la bioinformatique est devenue l'alliée indispensable des chercheurs. La bioinformatique est un champ multidisciplinaire qui utilise des méthodes informatiques, mathématiques, statistiques, physiques et combinatoires dans le but de visualiser, stocker, extraire, organiser et analyser les données biologiques obtenues par l'intervention des techniques de la biologie moléculaire et dont leurs interprétations mènera à de nouvelles connaissances (Xiong, 2006).

La bioinformatique est une branche théorique et pratique de la biologie. Sur le plan théorique, sa finalité est la synthèse des données biologiques à l'aide de modèles et de théories en énonçant des hypothèses généralisatrices et en formulant des prédictions. Sur le plan pratique, son but est de proposer des méthodes et des logiciels pour la sauvegarde, la gestion et le traitement de données biologiques. Pour cela, les Anglo-saxons, utilisent deux termes pour distinguer ces deux aspects de la bioinformatique: "bioinformatics" pour l'aspect pratique et "biocomputing" ou "computational biology" pour désigner l'aspect théorique (Keedwell and Narayanan, 2005).

Le terme « bioinformatique » date du début des années 80. Cependant, le concept sous-jacent de traitement de l'information biologique est bien plus vieux. Avant 1985, ce terme n'étais pas indexé comme mot clé par la base de données « Medline ». Jusqu'en 1992, il n'apparaît presque pas dans les titres ou les résumés référencés (une seul fois). En 1993, le terme apparaît enfin 6 fois puis 12 et 13 fois en $1994 - 95$ pour ensuite augmenter de façon exponentielle (Figure 2.5).

Les premiers articles dans le domaine ont le plus souvent été publiés dans les journaux « The journal of Molecular Biology », « Nucleic Acids Research » et « Computer Applications in Biological Sciences ». Ce dernier, fondé en 1985, devient en 1998 « Bioinformatics », qui devient aujourd'hui le journal de référence de la discipline. Désormais, plus d'une dizaine de journaux consacrés à la bio-informatique existent, tel que « BMC Bioinformatics », « BMC Genomics », « Computers in Biology and Medicine » etc .

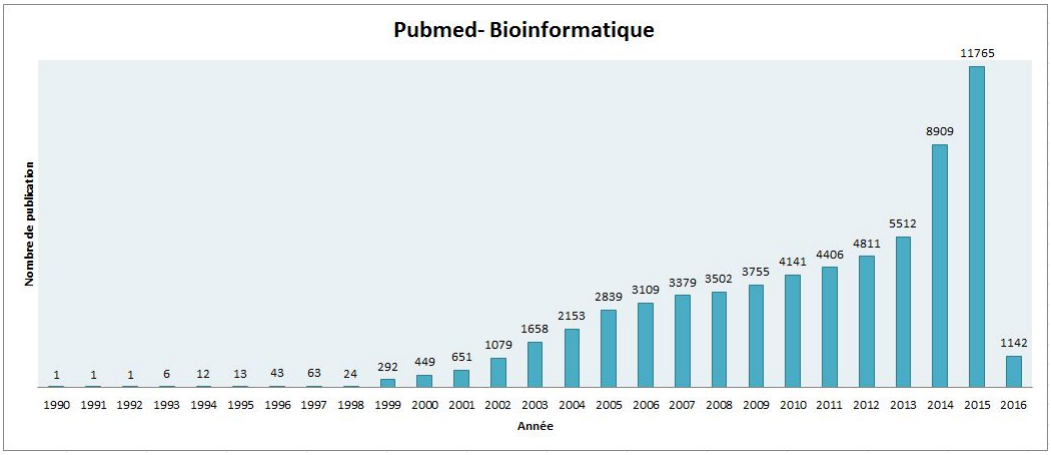


Figure 2.5: **La bio-informatique dans la littérature scientifique de 1990 au janvier 2016 (source : PubMed)**. Croissance exponentielle du nombre d'articles référencés dans PubMed sous le terme « bioinformatics ».

La bioinformatique va bien au-delà de la mode, elle est au centre des développements les plus récents de la biologie et la médecine, comme le décryptage du génome humain (un autre mot à la mode), les nouvelles techniques de biotechnologies et médico-légales (utilisées par la police scientifique), ainsi que la médecine personnalisée (Claverie and Notredame, 2011). De nos jours, la bioinformatique sert à :

- Mettre de l'ordre dans l'amoncellement des données biologiques en créant des banques de données.
- Analyser et comprendre les mécanismes de la vie en concevant des programmes bioinformatiques.
- Elaborer des concepts qui peuvent soutenir et aiguiller la recherche.
- Acquérir une vision nouvelle et plus globale des sciences de la vie.
- Modéliser des phénomènes biologiques comme par exemple la coagulation du sang ou la synthèse d'une protéine.
- Suggérer des prédictions sur la base de comparaison, telles que la fonction d'une protéine ou l'implication d'un gène dans une maladie.
- Créer de nouvelles synergies entre des scientifiques d'horizons différents, pour comprendre les mécanismes du vivant.
- Soutenir la recherche expérimentale en laboratoire, comme par exemple la recherche biomédicale pour le développement de nouveaux médicaments et de nouvelles thérapies.
- Découvrir des biomarqueurs pour aider au diagnostic des maladies et permettre la prédiction in Silico[2], qui est plus efficace et plus rapide que les techniques de diagnostic classiques (empiriques) d'une part et ouvre la porte pour une médecine personnalisée (thérapie ciblée) d'autre part.

2.3 Différent types de problèmes étudiés en bioinformatique

Les problèmes étudiés en bioinformatique actuellement sont (Valentini *et al.*, 2009; Mitra and Hayashi, 2006):

- La reconstruction et le séquençage du génome en entier (Lander *et al.*, 2001; Venter *et al.*, 2001).
- L'extraction et l'identification de la structure des gènes (Brent and Guigo, 2004; Bernal *et al.*, 2007).

[2]Signifie « Sur l'ordinateur », Utilisée pour indiquer que la recherche est effectuée dans l'ordinateur par opposition aux recherches dans un tube à essai ou dans le corps.

- L'identification et l'analyse des éléments d'ADN non-codants (Rätsch *et al.*, 2007; Holloway *et al.*, 2007).
- L'identification des gènes impliqués dans les maladies génétiques (López-Bigas and Ouzounis, 2004).
- La prévision des effets phénotypiques de polymorphisme nucléotidiques simples non synonymes (Bao and Cui, 2005).
- L'identification des éléments structuraux d'ARN (Bao and Cui, 2005).
- La modélisation des blocs d'haplotypes (Greenspan and Geiger, 2004).
- La prédiction d'un site d'épissage (Saeys *et al.*, 2004).
- La détection des interactions entre gènes, dans l'étude des maladies humaines (Ritchie *et al.*, 2003).
- L'alignement multiple de séquences biologiques (Handl *et al.*, 2007).
- La découverte des biomarqueurs biologiques génomiques (Wu *et al.*, 2012).
- La classification et le diagnostic du cancer à l'aide des données d'expression de gènes (Khan *et al.*, 2001; Lee and Lee, 2003; Kohlmann *et al.*, 2004).
- La reconstruction des réseaux de régulation de gènes à partir de données d'expression génique (Wang *et al.*, 2013).

2.4 Limites de la bioinformatique

Ayant reconnu les avantages et les point fort de la bioinformatique, il est également important de connaitre ses limites et éviter de trop compter sur seulement les outils bioinformatiques et leurs sorties. En fait, la bioinformatique a un certain nombre de limitations inhérentes. De plusieurs façons, le rôle de la bioinformatique en génomique et en recherche en biologie moléculaire peut être comparé au rôle de la collecte de renseignements dans les champs de bataille. Le renseignement est manifestement très important pour mener à la victoire dans un champ de bataille. Un combat sans intelligence est supérieure informations est inefficace et dangereux. Avoir une information supérieure et une intelligence correcte aide à identifier les faiblesses de l'ennemi et de révéler la stratégie et les intentions de l'ennemi. L'information recueillie peut ensuite être utilisée pour diriger les forces afin

d'engager l'ennemi et gagner la bataille. Cependant, compter complètement sur l'intelligence peut aussi être dangereux si l'intelligence est d'une précision limitée. Une dépendance excessive à une mauvaise qualité d'intelligence peut donner des erreurs coûteuses si on ne complète pas les défaillances (Xiong, 2006).

Par analogie, la lutte contre les maladies ou d'autres problèmes biologiques en utilisant la bioinformatique est comme des batailles avec de l'intelligence. La bioinformatique et la biologie expérimentale sont indépendants, mais complémentaires. La bioinformatique dépend de la science expérimentale pour produire des données brutes pour l'analyse. A son tour, elle fournit une interprétation utile des données et des pistes importantes pour la poursuite des recherches expérimentales. Les prédictions bioinformatiques ne sont pas des preuves formelles de tous les concepts. Ils ne remplacent pas les méthodes de recherche expérimentales traditionnelles. En outre, la qualité des prédictions bioinformatiques dépend de la qualité des données et la sophistication des algorithmes utilisés. Les données de séquence de l'analyse à haut débit contiennent souvent des erreurs. Si les séquences sont mal ou incorrectement annotées, ainsi les résultats de l'analyse en aval sont trompeurs. Voilà pourquoi il est si important de maintenir une perspective réaliste du rôle de la bioinformatique (Xiong, 2006).

La bioinformatique est en aucun cas un champ mature. La plupart des algorithmes n'ont pas la capacité et la sophistication de refléter réellement la réalité. Ils font souvent des prédictions erronées qui n'ont aucun sens lorsqu'ils sont placés dans un contexte biologique. Les erreurs dans l'alignement de séquence, par exemple, peuvent affecter le résultat de l'analyse structurelle ou phylogénétique. Le résultat du calcul dépend aussi de la puissance de calcul disponible. De nombreux algorithmes précis mais exhaustives ne peuvent pas être utilisés en raison de la complexité de calcul. Au lieu de cela, les algorithmes moins précis mais plus rapides doivent être utilisés. Ceci est un compromis nécessaire entre la précision et la faisabilité des calculs. Par conséquent, il est important de garder à l'esprit le risque d'erreurs produites par les outils bioinformatiques. La prudence devrait toujours être exercée lors de l'interprétation des résultats de prédiction. Il est une bonne pratique d'utiliser plusieurs outils, s'ils sont disponibles, et d'effectuer des évaluations multiples. Une prévision plus précise peut souvent être obtenue si l'on dessine un consensus en comparant les résultats de différents algorithmes.

2.5 Médecine personnalisée, biomarqueurs et cancer

2.6 Médecine personnalisée

Le concept de la médecine personnalisée date de plusieurs centaines d'années. Mais il a fallu attendre le 19ème siècle, que l'évolution de la chimie, l'histochimie et la microscopie ce qui a permis aux scientifiques de commencer à comprendre les causes sous-jacentes des maladies. Avec la croissance des industries pharmaceutiques et de dispositifs médicaux dans le 20ème siècle viennent l'apparition de la génétique, l'imagerie et l'extraction de données. Au milieu du siècle, les observations de différences individuelles dans la réponse aux médicaments ont donné lieu à une piste de recherche sur l'identification des enzymes clés qui jouent un rôle dans la variation du métabolisme des médicaments et la réponse, ce qui a servi de fondement à la pharmacogénétique.

Plus récemment, le séquençage du génome humain au début du 21ème siècle mis en mouvement la transformation de la médecine personnalisée à partir d'une idée à une pratique. L'évolution rapide de la génomique, ainsi que les progrès dans d'autres domaines, tels que la biologie computationnelle, l'imagerie médicale et la médecine régénérative, créent la possibilité pour les scientifiques à développer des outils afin de réellement personnaliser le diagnostic et le traitement.

Il n'y a aucune définition officiellement reconnue de la médecine personnalisée. Le terme «médecine personnalisée», a d'abord été utilisé comme le titre d'une monographie en 1998 (Jain, 2005) et a commencé à apparaître dans MEDLINE en 1999, mais la plupart de la littérature pertinente à la médecine personnalisée est toujours classée dans la pharmacogénomique et pharmacogénétique. Alors, La définition et la portée du terme médecine personnalisée varient considérablement, allant de la très large à la très étroite. Les exemples ci-dessous ont été choisis pour démontrer la portée des définitions qui ont été proposées:

- "L'utilisation de nouvelles méthodes d'analyse moléculaire pour mieux gérer la maladie ou la prédisposition d'un patient à la maladie." – "Personalized Medicine Coalition"
- "Offrir le bon traitement pour le bon patient, à la bonne dose au bon moment." – "European Union"

- "L'adaptation d'un traitement médical aux caractéristiques individuelles de chaque patient." – "President's Council of Advisors on Science and Technology"
- "Les soins de santé qui sont informés par l'information clinique, génétique et de l'environnement unique de chaque personne." – "American Medical Association"
- "Une forme de médecine qui utilise des informations sur les gènes, les protéines et l'environnement d'une personne pour prédire, diagnostiquer et traiter la maladie." – "National Cancer Institute, NIH"

D'autres part, la plupart des médicaments actuels sont approuvés et développés sur la base de leur performance dans un grand nombre de personnes, mais la médecine de l'avenir se développe comme une solution personnalisée pour les besoins d'un patient particulier. En cas de troubles complexes, les approches classiques "one-drogue-fits-all" impliquent plusieurs essais et erreurs avant qu'un traitement approprié est trouvé. Les données d'essais cliniques pour un nouveau médicament montrent simplement la réponse moyenne d'un groupe d'étude. Il ya, cependant, des variations individuelles considérables; certains patients ne présentent pas de réponse alors que d'autres présentent une réponse dramatique. Il est évident que le concept "un médicament pour tous les patients atteints de la même maladie" ne tient pas et une approche plus individualisée est nécessaire. Bien que l'individualisation de certains traitements ait été réalisée au niveau génomique, le concept de médecine personnalisée suit les progrès dans l'étude des maladies humaines au niveau moléculaire. Le but de la médecine personnalisée est de faire correspondre le bon médicament au bon patient et, dans certains cas, même pour concevoir le traitement pour un patient selon un génotype ainsi que d'autres caractéristiques individuelles. Un terme plus large est intégré "soins de santé (healthcare)", qui comprend le développement de la médecine fondée sur la génomique personnalisée, les tests de prédisposition, la médecine préventive, la combinaison de diagnostic avec le thérapeutique et le suivi de la thérapie (Jain, 2015).

2.7 Médecine conventionnelle Vs médecine personnalisée

Les médicaments conventionnels ont connu un début de thérapies empiriques. Alors, même que les thérapies basées mécanisme ont commencé

à se développer, le manque d'efficacité et les effets indésirables ont été observés et acceptés dans une certaine mesure. La plupart des médicaments conventionnels ont été développés comme médicaments universel pour une certaine maladie. Pour les maladies avec de multiples pharmacothérapies, le choix a été généralement laissé à l'expérience et aux préférences du médecin prescripteur. Avec les progrès de la pharmacogénétique, il est devenu évident que quelque chose pouvait être fait pour les problèmes suivants avec les médicaments conventionnels.

- Les variations génétiques entre les individus conduisent à des différences dans la réponse aux médicaments.
- Le pourcentage élevé de manque d'efficacité avec certains médicaments.
- La haute incidence des effets indésirables (secondaires) des médicaments.
- La médecine factuelle (Evidence-based medicine) supporte une application standardisée de thérapie qui ne prend pas en compte les variations de réponse chez les patients.
- Les essais cliniques sont axés autour la prise en compte des informations statistiques sur la population générale des patients et de l'appliquer à l'individu.

2.8 Qu'est-ce qu'un biomarqueurs?

Quand nous allons chez notre médecin pour une vérification annuel, nous sommes susceptibles d'avoir notre taux de cholestérol et la pression artérielle vérifiés. Ces procédures sont jugées importantes parce que l'hypercholestérolémie est un biomarqueur de la maladie cardiovasculaire et l'hypertension artérielle est un biomarqueur d'accident vasculaire cérébral. Autrefois, les médecins utilisent la couleur de l'urine de leurs patients pour déterminer s'ils étaient en bonne santé. Comme on peut le voir à partir de ces exemples, les biomarqueurs ont été avec nous depuis longtemps et sont devenus une partie de la routine des soins médicaux.

Idéalement, différentes organisations et publications seraient d'accord sur la définition d'un biomarqueur. Cependant, la définition des biomarqueurs n'est pas simple parce que le terme est utilisé dans un certain nombre de différentes disciplines et les types de mesures biologiques qui sont considérées

comme biomarqueurs ont élargi au fil du temps. Par exemple, la pression sanguine et du cholestérol démontrent l'utilisation de biomarqueurs en médecine. Cependant, les biomarqueurs sont également utilisés dans l'écologie pour indiquer la santé des écosystèmes ou les effets de l'intervention humaine sur les autres espèces animales. Dans ce qui suit, nous allons limiter notre discussion de biomarqueurs à ceux utilisés en médecine humaine et la recherche biomédicale. Même dans ces disciplines, ce qui est considéré comme un biomarqueur a changé au fil du temps que les nouvelles technologies ont été développées. Dans de nombreux domaines de la médecine, les biomarqueurs utilisés se limitaient à des protéines qui sont identifiables et mesurables dans le sang ou l'urine (Network, 2010).

Aujourd'hui, les techniques d'imagerie permettent de visualiser les aspects du corps que nous ne pouvions pas voir avant et ont abouti à la découverte de nombreux nouveaux biomarqueurs. Par exemple, des techniques d'imagerie permettent la détection des changements structurels dans le cerveau humain qui peuvent être utilisés en tant qu'indicateurs de certaines maladies ou conditions. À la suite de ces changements, la définition du terme biomarqueur nécessite un peu plus d'exploration.
Le Tableau 2.1 présente les définitions des biomarqueurs fournies par diverses organisations et publications. Comme on peut le voir dans ce tableau, la plupart des définitions de biomarqueurs sont constitués de deux parties:

- Quels genres de choses peuvent être des biomarqueurs?
- Quel est le but d'un biomarqueur? Cela dit, que faut-il indiquer ce biomarqueur?

Source	**Définition d'un biomarqueur**
« **National Cancer Institute** »	Une molécule biologique présente dans le sang, et d'autres fluides corporels ou des tissus qui est un signe d'un processus normal ou anormal, ou d'une condition ou d'une maladie. Un biomarqueur peut être utilisé pour voir comment le corps réagit à un traitement pour une maladie ou condition. Aussi appelé marqueur moléculaire ou signature moléculaire.
« **MedicineNet dictionary** »	Une caractéristique biochimique ou une facette qui peut être utilisée pour mesurer la progression de la maladie ou les effets du traitement.
« **Center for Biomarkers in Imaging (Massachusetts General Hospital)**»	Un paramètre anatomique, physiologique, biochimique ou moléculaire liés à la présence et la gravité des états pathologiques particuliers.
« **Biomarkers Consortium (Foundation of National Institutes of Health)**»	Les caractéristiques qui sont objectivement mesurées et évalués comme des indicateurs de processus biologiques normaux, des processus pathogéniques, ou des réponses pharmacologiques à une intervention thérapeutique.

Table 2.1: Différentes définitions d'un biomarqueur (Network, 2010)

2.9 Quel est l'objectif de l'utilisation des biomarqueurs?

La plupart des définitions de biomarqueurs ont noté qu'ils peuvent avoir au moins un des plusieurs objectifs suivant et qui peuvent être aussi des critères pour les classifier en (Network, 2010):

- **Biomarqueurs de diagnostic:** aident à diagnostiquer un cancer, peut-être avant qu'il ne soit détectable par les méthodes conventionnelles.
- **Biomarqueurs de pronostic:** prévoient le degré d'agressivité du processus de la maladie, et montrent comment un patient peut s'attendre à s'en sortir en l'absence de thérapie.
- **Biomarqueurs prédictifs:** aident à identifier quels patients répondront à quels médicaments et avec quel dose.

Le Tableau 2.2, présente des exemples de biomarqueurs pour chacun des types cités ci-dessus :

Stade	Maladie	Biomarqueur
Diagnostic	**Leucémie**	*PML-RARA* *BCR-ABL1* *CBFB-MYH11* *ETV6-RUNX1* *RUNX1-RUNX1T1* *MLL*-rearranged *TCF3-PBX1* *RBM15-MKL1*
	Sarcome	*SS18-SSX1/SSX2* *PAX3/ PAX7-FOXO1A* *EWSR1-FLI1* *EWSR1-ERG* *EWSR1-NR4A3* *TAF15-NR4A3* *EWSR1-ATF1* *EWSR1-CREB1* *ASPSCR1-TFE3* *FUS-DDIT3* *JAZF1-SUZ12*
Prédictive	**Cancer du sein**	*HER2*
Pronostic	**Cancer du sein**	*OncotypeDx* *Mammaprint* *IHC4*

Table 2.2: Exemples de biomarqueurs selon leurs objectifs (Gonzalez de Castro *et al.*, 2013)

2.10 Type de biomarqueurs

Les biomarqueurs utilisés aujourd'hui dans la médecine et la recherche se divisent généralement en plusieurs catégories. Des biomarqueurs moléculaires, également appelés marqueurs moléculaires ou les marqueurs biochimiques, sont l'un des types les plus courant et qui nous intéresse spécifiquement dans cette thèse. Ce sont souvent des gènes ou des protéines, telles que $HER-2/neu$ dans le cancer du sein. Cependant, comme nous l'avons vu, les processus physiologiques tels que la pression artérielle et la circulation sanguine sont également utilisés comme biomarqueurs, comme le sont également certaines structures anatomiques telles que la taille d'une zone du cerveau (Jain, 2010). Dans ce qui suit, nous décrivons ces trois catégories de biomarqueurs selon leurs natures, ainsi que quelques exemples, cités dans le Tableau 2.3.

- **Biomarqueurs moléculaires ou biochimiques:** sont des molécules biologiques trouvées dans les fluides corporels ou les tissus. Dans le cancer, les biomarqueurs moléculaires sont souvent des gènes, des ARN non-codants ou produits géniques tels que des protéines. Un exemple est l'antigène spécifique de la prostate qui est une protéine produite par les cellules de la prostate qui se trouve normalement dans de faibles niveaux dans le sang des hommes. Des niveaux accrus de l'antigène spécifique de la prostate sont utilisés en tant que biomarqueur de diagnostic pour le cancer de la prostate, bien que des niveaux élevés puissent également indiquer une inflammation de la prostate ou d'autres conditions.

- **Biomarqueurs physiologiques:** sont ceux qui ont à voir avec les processus fonctionnels dans le corps. Par exemple, le flux sanguin dans les zones cérébrales affectées est étudié comme un indicateur potentiel de succès du traitement. Comme les techniques d'imagerie deviennent plus avancées, nous sommes susceptibles de voir une augmentation de l'utilisation de biomarqueurs physiologiques.

- **Biomarqueurs anatomiques:** sont ceux qui ont à voir avec la structure d'un organisme et le rapport de ses parties. Les biomarqueurs anatomiques comprennent la structure de divers organes tels que le cerveau ou le foie. Par exemple, la taille de certaines structures du cerveau par rapport à d'autres est un biomarqueur pour un trouble du mouvement connu comme la maladie de Huntington. La découverte de biomarqueurs anatomiques est également facilitée par le développement des techniques d'imageries.

Biomarqueur	**Type**	**Condition**
Protéine C-réactive	Moléculaire/ biochimique	Inflammation
Taux de cholestérol élevé	Moléculaire/ biochimique	Maladie cardiovasculaire
Protéine S100	Moléculaire/ biochimique	Mélanome
Le gène HER-2/neu	Moléculaire/ biochimique	Cancer du sein
Les gènes BRCA	Moléculaire/ biochimique	Les cancers du sein et de l'ovaire
Antigène prostatique spécifique (PSA)	Moléculaire/ biochimique	Cancer de la prostate
CA-125	Moléculaire/ biochimique	Cancer de l'ovaire
Le débit sanguin cérébral	Physiologique	La maladie d'Alzheimer, accident vasculaire cérébral, la schizophrénie
Température du corps élevée	Physiologique	Infection
Taille des structures cérébrales	Anatomique	La maladie de Huntington

Table 2.3: Exemples de biomarqueurs selon leurs natures

2.11 Rôle de biomarqueurs en médecine personnalisée

Une des principales raisons de l'intérêt croissant dans les biomarqueurs, est le potentiel qu'ils détiennent pour la médecine individualisée ou personnalisée, aussi appelée thérapie ciblée. Certes, il existe une différence entre les types de cancers, ils ne sont pas tous pareils. De plus, en apprenant davantage sur les cellules cancéreuses et leur milieu environnant, le nombre des sous-types de chaque cancer augmente. Les sous-types sont souvent basés sur les biomarqueurs qui permettent de distinguer le cancer basée sur certaines caractéristiques importantes, telles que l'agressivité de la maladie (biomarqueur pronostic) ou la réponse au traitement (biomarqueur prédictif) (Azuaje, 2011).

La médecine personnalisée qui se concentre sur les différences entre les personnes et le potentiel de ce dernier d'influer les résultats médicaux. De sorte que, le cancer d'une personne peut être catégorisé selon certains biomarqueurs qui sont présents ou absents, augmentés ou diminués. Par conséquent, on aura une plus grande probabilité de recevoir un traitement approprié et efficace pour notre cancer particulier, contrairement à la méthode empirique utilisée dans le passé et actuellement pour déterminer le traitement.

Les méthodes bioinformatiques sont appliquées pour le développement et la validation de nouveaux biomarqueurs qui sont utiles pour choisir les bons traitements aux patients appropriés. L'hétérogénéité de la maladie établie sur la base de biomarqueurs génomiques nécessite le développement de nouveaux paradigmes de la conception et l'analyse des essais cliniques. Afin d'évaluer la validité et l'utilité clinique des nouveaux traitements et les biomarqueurs utilisés en médecine personnalisée (Matsui, 2013).

2.12 L'utilisation de biomarqueurs en cancer

Un biomarqueur peut avoir plus d'un emploi et certains biomarqueurs sont utilisés dans la médecine du cancer. Les biomarqueurs ont de nombreuses utilisations dans le cancer, non seulement dans le traitement des patients, mais aussi dans le développement de nouveaux médicaments. Donc, il existe deux principales utilisations des biomarqueurs et qui résument les deux grandes directions de recherches dans le cancer en bioinformatique citées ci-après.

2.12.1 L'utilisation de biomarqueurs en médecine du cancer

L'utilisation de biomarqueurs en médecine du cancer se résume dans (Gonzalez de Castro *et al.*, 2013; Cristofanilli *et al.*, 2004; O'Brien *et al.*, 2003; Terpos *et al.*, 2010; Ludwig and Weinstein, 2005):

- L'évaluation du risque d'avoir le cancer: Suis-je à risque accru de cancer?
- Le diagnostic: Est-ce que j'ai un cancer? Quel est le type du cancer dont je dispose?
- Pronostic: Quelle est l'évolution attendue de mon cancer, mais en l'absence du traitement?
- Prédiction de la réponse au traitement: Est-ce que le type du cancer que j'ai va répondre à ce médicament?
- Prédiction des doses de médicaments: Devrais-je recevoir une dose normale, inférieure ou non pas du tout utilisé ce médicament?
- Le suivi de la réponse au traitement: Comment mon cancer a répondu au traitement ?
- La récurrence du cancer: Est-ce que mon cancer va revenir ?

2.12.2 L'utilisation de biomarqueurs dans la découverte des médicaments

En plus de leur utilisation dans la médecine, les biomarqueurs sont également très couramment utilisés dans la découverte de nouveaux médicaments de cancer. Dans cette direction, il existe un grand défi qui est le développement des cibles appropriées de médicaments.

2.13 Les caractéristiques d'un biomarqueur idéal

Les facteurs que nous désirons idéalement trouver dans les tests de biomarqueurs du cancer sont (Network, 2010):

- Le biomarqueur est présent chez les personnes atteintes de la maladie, mais est très rarement présent chez les personnes sans la maladie.

- Le biomarqueur est présent dans un liquide corporel facilement accessible tel que l'urine.
- Le biomarqueur est facilement détecté, par un test standardisé fiable, valide et simple à réaliser.

2.13.1 Spécificité et sensibilité de biomarqueurs

Ces deux concepts sont des déterminants majeurs de l'utilité d'un biomarqueur ou un test de biomarqueur, ils sont très importants à considérer dans les outils bioinformatiques proposés pour les découvrir.

- Spécificité: La probabilité d'obtenir un résultat négatif lorsque la cible (le biomarqueur) n'est pas présente.
- Sensibilité: la probabilité d'obtenir un résultat positif lorsque la cible (le biomarqueur) est effectivement présente.

2.14 Les méthodes de découverte de biomarqueurs

On peut classer les méthodes de découverte de biomarqueurs selon l'unité biochimique à étudier, en quatre types (Nagaraj, 2009):

- Génomiques: science qui étudie le génome ou l'ensemble de l'ADN d'un organisme et ainsi que sa fonction (Zhang *et al.*, 2011a).
- Transcriptomiques: science qui étudie tous les gènes qui sont transcrits en ARN dans une cellule ou un organisme à un instant donné (Zhang *et al.*, 2011a).
- Protéomiques: science qui étudie toutes les protéines exprimées dans une cellule ou un organisme à un point de temps donné (Srinivas *et al.*, 2002).
- Métaboliques: science qui étudie tous les métabolites dans une cellule, tissu ou un organisme dans des conditions données (Kim *et al.*, 2008).

Dans le cadre de cette thèse, nous nous positionnons sur les méthodes de découverte de biomarqueurs génomiques et transcriptomiques, en fixant comme objectif l'exploration et l'utilisation des méthodes de l'intelligence computationnelle, afin de pouvoir trouver une meilleure solution pour découvrir ces biomarqueurs.

2.15 Méthodes génomiques de découverte de biomarqueurs

La génomique est l'étude multiples des gènes et la façon dont ils travaillent ensemble. Les méthodes génomiques sont utilisées pour déterminer quels sont les gènes qui différencient les cellules cancéreuses des cellules normales. Ces gènes peuvent être considérés comme des biomarqueurs. En effet, l'une des questions importantes que les chercheurs doivent se posée est où dans le génome nous devons regarder. Etant donné que, les êtres humains possèdent des milliers de gènes, il est difficile de comprendre quels sont les gènes qui sont les plus susceptibles d'être des biomarqueurs (Murray *et al.*, 2007). Ces méthodes se distinguent en deux grandes catégories (Ghosh and Chinnaiyan, 2005):

2.15.1 Approches globales/exhaustives (comprehensive approach)

Dans cette catégorie de méthodes, la découverte de biomarqueurs se fait sans le biais d'hypothèses scientifiques antérieures. Par ailleurs, les chercheurs analysent l'ensemble complet de l'ADN, et essayent de rapporter un profil d'expression génique dans une certaine caractéristique du cancer.

- Exemple: Balayer l'ensemble des gènes exprimés dans un groupe d'individus atteints d'un cancer du pancréas, et les comparer avec ceux d'un ensemble de gènes exprimés dans un groupe d'individus sans cancer du pancréas.

2.15.2 Approches fondées sur une hypothèse (the candidate-driven or hypothesis-driven approaches)

Dans cette approche, les chercheurs commencent par déterminer les gènes qui doivent etre examinés sur la base de la littérature scientifique préexistante.

- Exemple: Comparer l'expression de gènes impliqués dans la présélection d'une croissance cellulaire, dans un groupe d'individus atteints d'un cancer du pancréas, à celles d'un groupe d'individus sans cancer du pancréas.

Chacune de ces deux méthodes a ses avantages et ses inconvénients, comme illustré dans le Tableau 2.4.

/	**Approches fondées sur une hypothèse**	**Approches globales/exhaustives**
Avantage	- Met l'accent sur les voies ou les gènes qui ont plus forte probabilité d'être candidats retenus.	- Impartiale, moins susceptibles de manquer des gènes ou des voies principales / importantes.
Inconvénient	-Approche partiale qui peut manquer des gènes / voies importantes. -Dépend fortement de la base de connaissances existante, qui est souvent limitée et incomplète	- Nécessite une grande quantité de données ou un grand nombre de patients pour établir la puissance statistique raisonnable (nécessaire). - Forte probabilité de fausses associations entre les gènes. - La solution du problème est très complexe. - Grande quantité de données peut être irrésistible.

Table 2.4: Avantages et inconvénients des méthodes génomiques de découverte de biomarqueurs

CHAPITRE 3

Intelligence computationnelle en bioinformatique

Contenu du chapitre

3.1 Introduction

Au cours des dernières décennies, nous avons connu une croissance massive de l'information biologique recueillies par les communautés scientifiques connexes. Le traitement de grandes quantités de données délivrées par les bio-technologies à haut débit nécessite des procédures de gestion de données avancées d'une part pour un stockage efficace et la récupération de l'information biologique. D'autre part, il nécessite des méthodes raffinées pour extraire et analyser des connaissances biologiques à partir de ces données (Goble and Stevens, 2008). Les méthodes d'intelligence computationnelle (IC) et d'apprentissage automatique (AA) sont largement utilisées pour l'extraction de connaissances biologiques à partir de données biomoléculaires (Hassanien *et al.*, 2013). Afin d'obtenir des modèles à la fois pour représenter la connaissance biologique et prédire les caractéristiques des systèmes biologiques. Il est à noter qu'un nombre croissant de méthodes de IC et AA a été appliqué et souvent mis au point pour traiter un large éventail de problèmes de bioinformatique en génomique, transcriptomique, protéomique, l'analyse de l'expression de gènes, l'évolution biologique, la biologie des systèmes, et d'autres domaines pertinents en bioinformatique.

L'IC est un successeur de l'intelligence artificielle et qui signifie « Plusieurs choses pour plusieurs personnes ». En informatique c'est la capacité d'une

machine à réagir dans un environnement de différentes manières, et qui prend des décisions utiles en fonction des informations courantes. L'IC est utilisée pour résoudre des problèmes difficiles qui ne peuvent pas être résolus par les techniques classiques de l'intelligence artificielle (Kahraman *et al.*, 2010). Les composants d'IC doivent avoir les caractéristiques suivantes:

- Un potentiel considérable pour résoudre des problèmes du monde réel.
- La possibilité d'apprendre à partir de nouvelles expériences.
- La capacité de l'auto organisation.
- La possibilité d'adaptation pour répondre aux changements des conditions et de contraintes.

Sans doute, l'IC comprend de ces paradigmes dans l'intelligence artificielle qui se rapportent à une sorte de système biologique ou d'origine naturelle. Un consensus général suggère que ces paradigmes sont les réseaux de neurones, les algorithmes évolutionnaires, l'intelligence en essaim, et les systèmes flous (voir Figure 3.1) (Sumathi and Paneerselvam, 2010). Les réseaux de neurones sont basés sur leurs homologues biologiques dans le système nerveux humain. De même, l'informatique évolutive appuie fortement sur les principes de l'évolution darwinienne observée dans la nature. L'intelligence en essaim, à son tour, est calquée du comportement social des insectes et la chorégraphie des oiseaux grégaires. Enfin, le raisonnement humain en utilisant des termes linguistiques, imprécis, ou flous est approché par les systèmes flous (Zurada *et al.*, 1995).

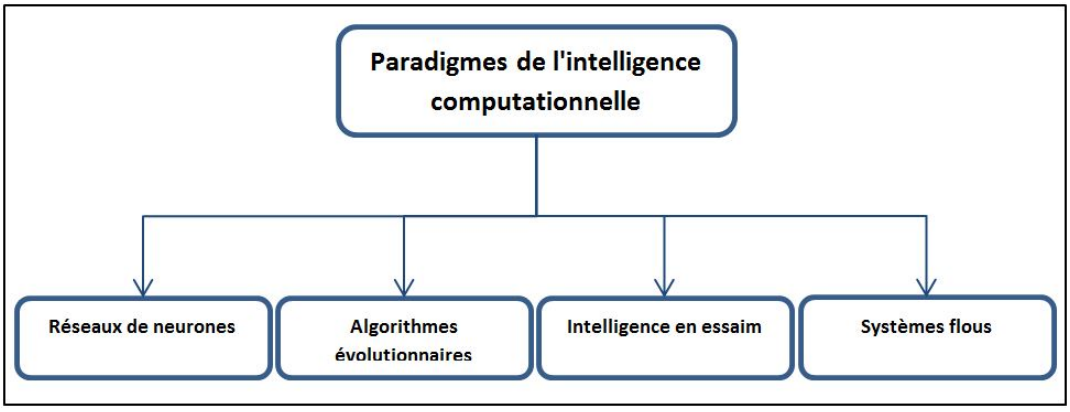


Figure 3.1: Paradigmes de l'intelligence computationnelle

Récemment, nous avons constaté une nouvelle époque émergente de l'IC qui met l'accent sur les principes, les aspects théoriques et la méthodologie de conception d'algorithmes inspirés de la nature. Par exemples, les réseaux de neurones artificiels inspirés des systèmes de neurones de mammifères, le calcul évolutif inspiré de la sélection naturelle en biologie, le recuit simulé inspiré des principes de la thermodynamique, et l'intelligence en essaim inspirée du comportement collectif des insectes ou des micro-organismes qui interagissent localement avec leur environnement causant une tendance globale, fonctionnelle et cohérente pour émerger (Hassanien *et al.*, 2013).
Les principales classes de problèmes en IC sont regroupées en cinq catégories qui sont: les problèmes de configuration, les problèmes d'optimisation, les problèmes de classification, les problèmes de régression, et les problèmes NP complets (Non-Deterministic Polynomial Time) (Sumathi and Paneerselvam, 2010). Dans cette thèse, nous abordons principalement des problèmes de classification supervisée, y compris la sélection de caractéristiques et d'optimisation.

3.2 Classification supervisée

Vu que la quantité et la variété des données disponibles augmentent, il se pose un besoin proportionnel pour des techniques d'exploration de données robustes, efficaces et polyvalentes qui peuvent être supervisées ou non supervisées. La classification supervisée est un sujet clé dans la discipline de l'apprentissage automatique et sert à attribuer des étiquettes de classe à un ensemble d'objets sous la supervision d'un enseignant (model de classification) (Mitra and Acharya, 2005).
Des frontières de décision sont générées pour discriminer des objets appartenant à des classes différentes. Ces objets sont d'abord divisés en un ensemble d'apprentissage et un autre de test. L'ensemble de test est utilisé pour évaluer la capacité de généralisation du classificateur. Chacun de ces deux ensemble est constitué d'un nombre N d'échantillons indépendants. Chaque échantillon (exemple) est un couple $(x; u)$, ou $x \in X$ est la description ou la représentation d'un groupe de variables ou de caractéristiques $x_1...x_M$ et $u \in U$ représente la supervision de x. Dans un problème de classification, u s'appelle la classe de x qui «supervise» le processus en cours et appartient à un ensemble $C : u_1....u_C$. C désigne le nombre de classes possibles pour un objet. C doit être également fini et en pratique petit pour que l'on puisse réellement parler de classification (Cornuéjols and Miclet, 2011).
Les valeurs de classes d'exemples peuvent être numériques ou de nature catégorique. Une valeur de classe numérique a des valeurs continues et quanti-

tatives. D'autre part, un attribut catégorique prend des valeurs symboliques discrètes qui peuvent aussi être des étiquettes ou des catégories de classe. Si les valeurs de classes sont catégoriques, le problème est appelé classification avec cet attribut étant appelé l'étiquette de classe. Par contre, s'ils sont numériques, le problème est appelé régression. L'objectif de la classification et la régression est de construire un modèle concis de la distribution des attributs de classes. Une fois les étapes de prétraitement nécessaires sont effectuées sur les données disponibles, un algorithme de classification supervisée utilise la base de données d'entrainement pour induire un classificateur. Le modèle qui en résulte est utilisé pour attribuer des valeurs à une base de données de test, où les valeurs des attributs de prédiction sont connues, mais l'attribut de classe est à déterminer (voir Figure 3.2).

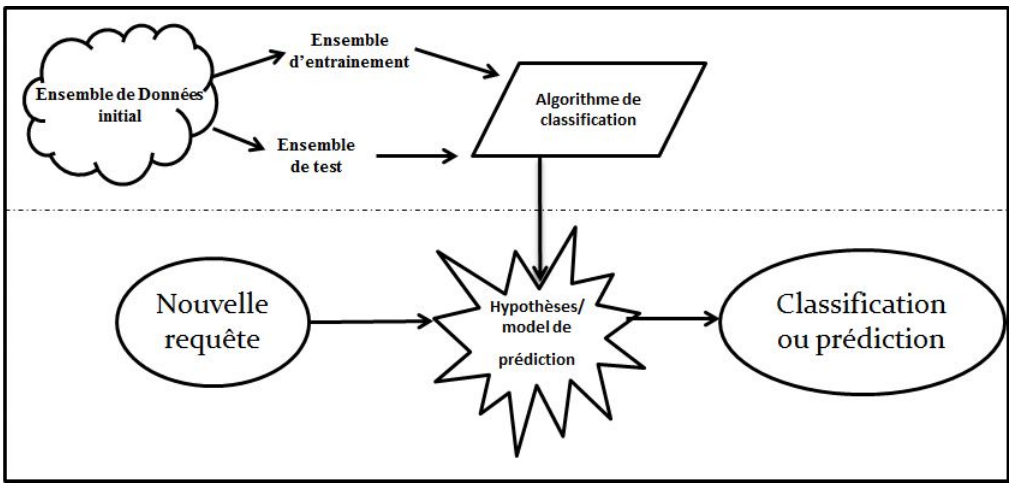


Figure 3.2: Principe de la classification supervisée

La classification supervisée est largement utilisée pour résoudre des problèmes très différents en bioinformatique tels que la prédiction de structure secondaire des protéines, le diagnostic basé sur l'expression de gènes, ou la prédiction de sites d'épissage. Les techniques de classification supervisée actuelles ont été montrées capable d'obtenir des résultats satisfaisants (Mitra and Acharya, 2005). Nous présentons ci-dessous les principales caractéristiques des techniques et modèles de classification les plus connus.

3.2.1 Machines à vecteurs de support (SVM)

Les techniques d'apprentissage basées noyau (comme SVM, Bayes, analyse en composantes principales, et les processus de Gauss) représentent une entité

majeure dans l'apprentissage automatique et les algorithmes d'intelligence computationnelle. Les SVMs ont d'abord été suggérés par Vapnik dans les années 1960 pour la classification et ont récemment devenu un domaine de recherche intense en raison de l'évolution des techniques et de la théorie couplée avec des extensions à la régression et l'estimation de densité (voir Figure 3.3)(Vapnik, 1998).

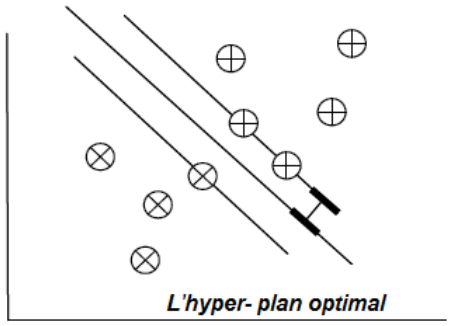


Figure 3.3: Machines à vecteurs de support

Les SVMs sont un groupe de méthodes d'apprentissage supervisé qui peuvent être appliqués à la classification ou à la régression. La classification est réalisée par une surface de séparation linéaire ou non linéaire dans l'espace de l'ensemble de données d'entré. SVM donne de bons performances dans plusieurs applications du monde réel tels que la catégorisation de textes, la reconnaissance de caractères écrits à la main, la classification d'images, l'analyse bio-séquences, etc., et il est maintenant établi comme l'un des outils standard pour IC et l'exploration de données (Hassanien *et al.*, 2013). SVM emploie des noyaux pour mapper les données d'entrées dans un espace de caractéristiques d'une dimension plus élevée implicitement et dans lequel les données deviennent linéairement séparables. La frontière de décision linéaire est établie de façon que la marge (distance minimale entre les exemples d'apprentissage et la limite) soit maximisée. Dans le cas où les points de données mappées sont linéairement inséparables, un coût est inclus pour tenir compte des exemples mal classés et la marge est maximisée en minimisant ainsi le coût (Burges, 1998).

3.2.2 Réseaux de neurones artificiels (ANN)

Les réseaux de neurones artificiels (ANN) ont été développés comme des généralisations des modèles mathématiques de systèmes nerveux biologiques.

Dans un modèle mathématique simplifié d'un neurone, les effets des synapses sont représentés par des poids de connexion qui modulent l'effet des signaux d'entrée associés, tandis que la caractéristique non linéaire présentée par les neurones est représentée par une fonction de transfert (voir Figure 3.4) (Haykin, 1999).

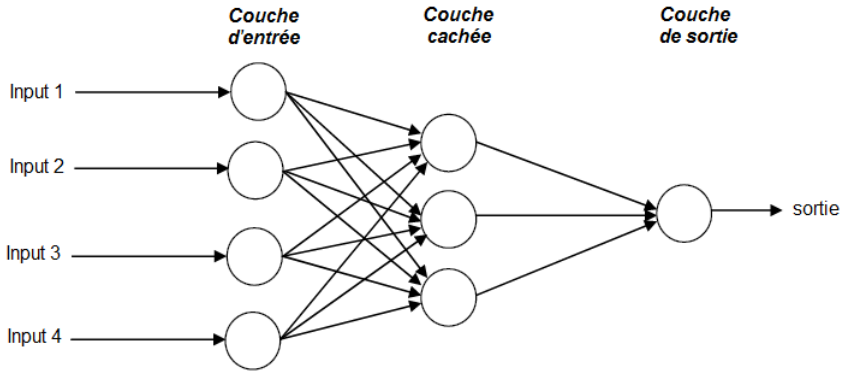


Figure 3.4: Réseau de neurones artificiel

Chaque neurone est caractérisé par un niveau d'activité (ce qui représente l'état de polarisation d'un neurone), une valeur de sortie (représentant le taux d'allumage du neurone), un ensemble de connexions d'entrée (représentant les synapses sur la cellule et son dendrite), une valeur de polarisation (représentant un niveau de repos interne du neurone), et un ensemble de connexions de sortie (représentant les projections axonales de neurones) (Bishop, 1995). Chacun de ces aspects est représenté mathématiquement par des nombres réels. Ainsi, chaque connexion a un poids associé (force synaptique), qui détermine l'effet de l'entrée arrivée sur le niveau d'activation de l'unité. L'impulsion du neurone est donc calculée comme la somme pondérée des signaux d'entrée, transformée par la fonction de transfert. La capacité d'apprentissage d'un neurone artificiel est obtenue en ajustant les coefficients de pondération selon un algorithme d'apprentissage choisi (Hassanien *et al.*, 2013).

3.2.3 Bayésien naïf (NB)

Cette famille de classificateurs offre un large éventail de possibilités pour modéliser $P(c|x_1, x_2, ..., x_n)$, qui est le terme de probabilité de distribution de classe conditionné à chaque valeur possible des variables prédictives. Ce terme, en conjonction avec la probabilité a priori de la classe $P(c)$ et au

moyen de la règle de Bayes, est utilisé pour attribuer la plus probable classe posteriori à un nouvel échantillon invisible:

$$\gamma(x) = arg \quad max_c \quad P(c|x_1, x_2, ..., x_d) = arg \quad max_c \quad P(c) \quad P(x_1, x_2, ..., x_d|c) \tag{3.1}$$

Tous les paramètres statistiques sont calculés à partir des données d'entrainement, souvent par leurs estimateurs de la maximum vraisemblance. Selon le degré de complexité des relations entre les variables du problème à modéliser, de nombreux classificateurs bayésiens intéressants peuvent être trouvés dans la littérature.

Naïve Bayes est le classificateur le plus populaire des classificateurs bayésiens. Cela suppose que toutes les variables de domaine sont indépendantes lorsque la valeur de la classe est connue. Cette hypothèse simplifie considérablement les statistiques exposées, et seul les termes de classe conditionnées univariées $p(x_i|c)$ sont nécessaires. Bien que cette hypothèse est clairement violée dans de nombreuses situations (en particulier dans de nombreux problèmes réels avec une complexité inhérente), le classificateur de Bayes naïf est en mesure d'obtenir suffisamment de résultats précis dans de nombreux cas (Matthiesen, 2010).

3.2.4 k-plus proche voisins (KNN)

L'idée de base de l'algorithme de k-plus proche voisins est de classifier un échantillon non étiqueté en l'affectant à la classe la plus fréquente parmi ses k échantillons les plus proches (Figure 3.5). Bien que beaucoup de méthodes dans cette direction aient été proposées, la technique du vote majoritaire parmi les k échantillons les plus proches est la plus couramment utilisée. D'autres variantes ont été également proposées à savoir la distance pondérée plus proche voisin ("distance weighted nearest-neighbor") et "nearest-hyperrectangle".
Les implémentations de ces algorithmes utilisent couramment la distance euclidienne pour les attributs numériques et le chevauchement nominale pour les caractéristiques symboliques. D'autres mesures de distance utilisées incluent, la distance de "Mahalanobis" et "the modified value difference" (MVDM) pour les caractéristiques numériques et symboliques, respectivement (Witten and Frank, 2005).

La technique des KNN aussi connue sous le nom de "apprentissage par exemple" ou "apprentissage paresseux", ne provoque pas une expression explicite du modèle de prédiction. Bien que la technique de KNN soit en mesure

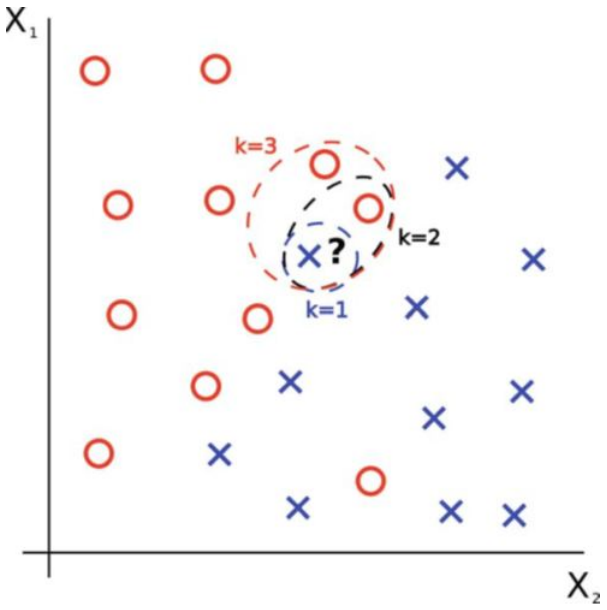


Figure 3.5: Exemple de classification avec KNN

d'obtenir des précisions de prédiction compétitive dans de nombreux problèmes, cette technique n'est pas utilisée dans de nombreuses situations réelles où une sortie de découverte de connaissances descriptive est nécessaire. Cela est dû à l'absence d'un modèle explicite à vérifier et à observer par des experts du domaine (Matthiesen, 2010).

3.2.5 Apprentissage à noyaux multiples (MKL)

Au cours des dernières années, plusieurs méthodes ont été proposées pour combiner plusieurs noyaux au lieu d'utiliser un seul. Ces différents noyaux peuvent correspondre à l'aide de différentes notions de similarité ou en utilisant des informations provenant de sources multiples (différentes représentations ou différents sous-ensembles de caractéristique). La fusion de plusieurs noyaux et les machines à vecteurs de support (SVM) (Vapnik, 2013), est une méthode d'apprentissage automatique supervisée bien connue qui a été largement utilisée dans divers domaines de la bioinformatique.
L'SVM est l'un des classificateurs basé noyau, qui peut trouver des limites non linéaires entre les classes de données à l'aide de noyaux. L'utilisation des méthodes basées noyau permet de représenter les données d'origine en utilisant une représentation matricielle, appelée matrice du noyau. Les matrices symétriques positives du noyau codent la similitude entre des échantillons (séquences) dans leur espace d'entrée respective. Cela implique que les car-

actéristiques hétérogènes peuvent toutes être remplacées par des matrices noyau de façon appropriée. Ceci permet l'élimination de l'hétérogénéité des données. La construction d'une même représentation pour tous les ensembles de données et l'intégration de ces représentations est l'intuition principale derrière les méthodes de fusion des noyaux.

Au cours des dernières années, plusieurs méthodes d'apprentissage automatique ont été proposées pour exploiter différentes sources d'information à l'aide de noyaux (Gönen and Alpaydın, 2011). Dans un tel cas, une bonne procédure pour la combinaison des noyaux implique une bonne combinaison des entrées provenant des sources multiples. Les recherches dans l'apprentissage à noyau multiple (MKL) sont concentrées à la fois sur le développement de nouvelles formulations ainsi que les optimiser. Différentes formulations sont nécessaires pour répondre aux besoins des différentes applications. La plupart des méthodes utilisant ces formulations proposent d'apprendre les noyaux combinés en réglant automatiquement les poids de chaque noyau (Gönen and Alpaydın, 2011). Les premiers travaux ont porté sur l'apprentissage du MKL comme une combinaison linéaire des noyaux de base (Lanckriet *et al.*, 2004).
Les combinaisons non linéaires de noyaux (Cortes *et al.*, 2009), comme les produits de noyaux et le mélange de polynômes, ont également été montrées pour être appropriées dans certains domaines. Beaucoup de ces formulations peuvent être facilement coulées dans le cadre généralisé du MKL (GMKL) proposé dans (Varma and Babu, 2009). Ce dernier a été utilisé dans l'une des contributions proposées dans cette thèse pour la prédiction des ARNnc, dans un cadre d'une étude intégrative à partir de plusieurs sources de données représentant plusieurs caractéristiques de ces molécules. Donc, l'apprentissage à noyaux multiples est utile dans la pratique et il y'a suffisamment de preuves indiquant que de meilleurs algorithmes MKL peuvent être conçus pour améliorer la précision et diminuer la complexité et le temps d'apprentissage.

3.3 Sélection de caractéristiques

3.3.1 Description informelle du problème

Dans une époque où la complexité et le volume des données disponibles dans le domaine de l'apprentissage automatique augmente chaque jour, la sélection de caractéristiques est devenue un élément indispensable du processus d'apprentissage(Bolón-Canedo *et al.*, 2014b). Elle se développe rapide-

ment à la fois en profondeur et en largeur, en raison de la demande croissante de la réduction de la dimensionnalité dans plusieurs applications. La sélection de caractéristiques est une étape de prétraitement importante dans de nombreuses applications d'apprentissage automatique, y compris la bioinformatique et la biologie computationnelle, où elle est généralement utilisée pour trouver le plus petit sous-ensemble de caractéristiques qui augmente extrêmement les performances du modèle de classification utilisé (Okun and Skarlas, 2011).

Par exemple, dans les ensembles de données d'expression génique, le nombre de gènes est beaucoup plus grand que le nombre d'échantillons. Cependant, il y a un grand nombre de gènes inutiles, redondants, ou bruyants. La présence de nombreuses caractéristiques affecte non seulement les performances de prédiction, mais aussi le temps d'exécution des algorithmes d'apprentissage. C'est pourquoi, nous avons besoin de techniques de réduction de dimensions, qui permettent d'identifier un petit ensemble de gènes, qui représente l'information la plus discriminante de l'ensemble de caractéristiques d'origine (initial), l'intégration de cette étape permet de (Boln-Canedo *et al.*, 2016):

- Améliorer de manière significative l'intelligibilité du classificateur, et maximiser les performances de prédiction d'algorithme de classification, dans le cas de la classification supervisée, et obtenir une meilleure détection des clusters en cas du regroupement (clustering),
- Réduire le coût d'acquisition et de stockage, face à la dégradation des performances de classification en raison de la finitude du nombre d'échantillons,
- Eviter le sur-apprentissage et réduire le temps d'apprentissage et de prédiction,
- Faciliter la visualisation et la compréhension de données.

Rajouter à tout cela, dans le scénario actuel de l'analyse des Big Data, la sélection de caractéristiques joue un rôle central.

Contrairement à d'autres techniques de réduction de dimensionnalité, comme celles basées sur la projection et la transformation de données (par exemple, l'analyse en composantes principales (Lu *et al.*, 2007; Sweilam *et al.*, 2010)) ou sur la compression (par exemple, la théorie de l'information), les techniques de sélection de caractéristiques ne modifient pas la représentation originale des variables. Mais simplement, elles sélectionnent un sous-ensemble

d'entre ces variables, qui représente dans notre cas des biomarqueurs. Ces derniers sont les gènes qui sont capables de faire la différence entre des échantillons provenant de différents populations ou de manière plus générale, les gènes qui sont pertinents et discriminants pour une annotation de cible particulière, cela sans transformer l'ensemble de gènes original (Yang *et al.*, 2010). La sélection de caractéristiques peut être réalisée de deux manières: La première consiste à classer les caractéristiques en fonction de certains critères et sélectionner les meilleurs k caractéristiques, et l'autre est de choisir un sous-ensemble minimal de caractéristiques sans avoir à apprendre de dégradation des performances. En d'autres termes, les algorithmes de sélection en sous-ensemble peuvent automatiquement déterminer le nombre de caractéristiques à sélectionner, tandis que les algorithmes de classement doivent pouvoir compter sur un certain seuil donné pour sélectionner les caractéristiques (Liu and Motoda, 2007). Plus de détails sur la représentation des solutions dans les algorithmes de sélection de caractéristiques sont donnés dans le chapitre 4. Étant donné le nombre croissant des jeux de données de grande dimension, les algorithmes de sélection de caractéristiques ont obtenus un intérêt croissant dans le domaine de l'apprentissage automatique, pour traiter beaucoup de ses disciplines, telles que le clustering, la régression et la classification, par les deux manières supervisées ou non supervisées ou plus récemment semi-supervisées (Boln-Canedo *et al.*, 2016). Dans le cadre de cette thèse, nous nous intéressons aux méthodes de sélection de caractéristiques supervisées, dans le sens où les données utilisées ont des étiquettes de classe. Par conséquent, les algorithmes de sélection supervisés reposent sur des mesures qui tiennent compte de l'information de classe.

3.3.2 Schéma général de la sélection de caractéristiques

Un processus de sélection de caractéristique typique se compose de quatre étapes de base (Figure 3.6), à savoir, la génération du sous-ensemble, l'évaluation du sous-ensemble généré, un critère d'arrêt, et la validation du sous-ensemble final (Dash and Liu, 1997). La génération d'un sous-ensemble est une procédure de recherche qui produit des sous-ensembles de caractéristiques candidats à être évaluer, basée sur une stratégie de recherche spécifique. Chaque sous-ensemble candidat est évalué et comparé avec le meilleur précédent selon un critère d'évaluation. Si le nouveau sous-ensemble se révèle être mieux, il remplace le meilleur sous-ensemble précédent. Le procédé de génération et d'évaluation de sous-ensembles est répété jusqu'à ce qu'un critère d'arrêt donné soit satisfait. Ensuite, le meilleur sous-ensemble sélectionné doit généralement être validé par des connaissances préalables ou par un classificateur via différents tests sur des ensembles de données synthétiques

et/ou réels (Liu and Yu, 2005).

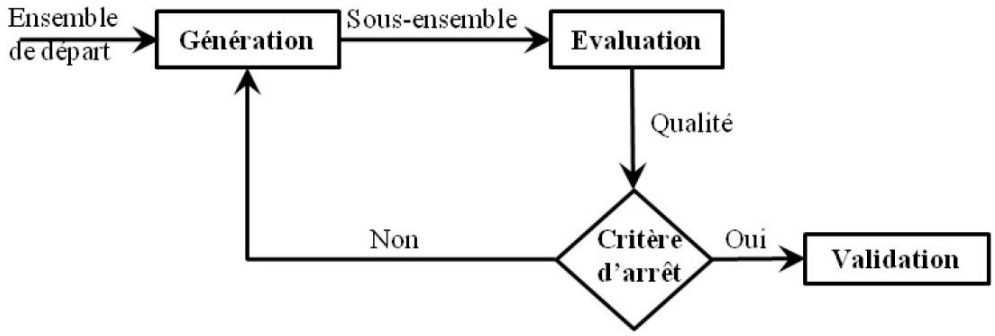


Figure 3.6: Schéma général d'un algorithme de sélection de caractéristiques

Génération

La génération des sous-ensembles est essentiellement un processus de recherche par heuristique, en spécifiant un sous-ensemble candidat pour l'évaluation à chaque état dans l'espace de recherche. La nature de ce processus est déterminée par deux questions fondamentales. Tout d'abord, on doit décider le point de départ de la recherche (un ou plusieurs points) qui à son tour influence la direction de recherche. La recherche peut commencer avec un ensemble vide et ajouter successivement des caractéristiques (vers l'avant ("forward")), ou de commencer avec un ensemble complet et retirer successivement des caractéristiques (vers l'arrière ("backward")), ou commencer par les deux extrémités et ajouter et supprimer des caractéristiques simultanément (bidirectionnel). La recherche, peut aussi commencer par un sous-ensemble sélectionné de façon aléatoire afin d'éviter d'être pris au piège, dans un optimum local (Doak, 1992). Deuxièmement, il faut décider une stratégie de recherche. Pour un ensemble de données avec N caractéristiques, il existe 2^N sous-ensembles de candidats. Cet espace de recherche est exponentiellement prohibitif pour la recherche exhaustive, même avec un N modéré. Par conséquent, différentes stratégies ont été explorées: recherche complète, séquentielle et aléatoire.

Recherche complète garantit de trouver le résultat optimal en fonction du critère d'évaluation utilisé. Bien qu'une recherche exhaustive est complète (par exemple, aucun sous-ensemble optimal est manqué), une recherche n'a pas à être exhaustive afin de garantir la complétude. Des fonctions heuristiques différentes peuvent être utilisées pour réduire l'espace de recherche

sans mettre en péril la possibilité de trouver un résultat optimal. Par conséquent, bien que l'ordre de l'espace de recherche est de $O(2^N)$, seulement un plus petit nombre de sous-ensembles seront évalués. On peut citer comme exemples d'heuristiques « branch and bound » (Narendra and Fukunaga, 1977), et la recherche en faisceau (beam search) (Doak, 1992).

Recherche séquentielle abandonne l'exhaustivité et risque ainsi de perdre des sous-ensembles optimaux. Il existe de nombreuses variantes de l'approche "hill-climbing" gourmande, comme la sélection séquentielle vers l'avant, élimination séquentielle vers l'arrière, et la sélection bidirectionnelle (Liu and Motoda, 2012). Toutes ces approches ajoutent ou suppriment des caractéristiques (une à la fois). Une autre alternative est d'ajouter (ou supprimer) p caractéristiques en une seule étape et de supprimer (ou ajouter) q caractéristiques dans la prochaine étape ($p > q$) (Doak, 1992). Les algorithmes avec une procédure de recherche séquentielle sont simples à mettre en œuvre et rapide pour produire des résultats, et que l'ordre de l'espace de recherche est généralement $O(N^2)$ ou moins.

Recherche aléatoire commence par un sous-ensemble sélectionné de façon aléatoire et procède de deux façons différentes. La première consiste à suivre la recherche séquentielle, qui intègre le caractère aléatoire dans les approches séquentielles classiques ci-dessus. Des exemples sont "random-start hill-climbing" et le recuit simulé (Doak, 1992). L'autre façon est de générer le prochain sous-ensemble d'une manière complètement aléatoire (par exemple, un sous-ensemble actuel n'est pas mise à jour à partir de tout sous-ensemble précédent suite à une règle déterministe), aussi connu comme l'algorithme "Las Vegas" (Brassard and Bratley, 1996). Pour toutes ces approches, l'utilisation de l'aspect aléatoire aide à échapper aux optima locaux dans l'espace de recherche, et l'optimalité du sous-ensemble sélectionné dépend des ressources disponibles.

Evaluation

Comme nous l'avons mentionné précédemment, chaque sous-ensemble nouvellement généré doit être évalué par un critère d'évaluation. La qualité d'un sous-ensemble est toujours déterminée par un certain critère (par exemple, un sous-ensemble optimal sélectionné en utilisant un critère peut ne pas être optimal selon un autre critère). Un critère d'évaluation peut être classé en deux groupes en fonction de leur dépendance à des algorithmes de classification qui sera finalement appliqué sur le sous-ensemble de caractéristiques sélectionné.

Critères indépendants d'un classificateur En général, un critère indépendant est utilisé dans les algorithmes du modèle de filtre. Ils essaient d'évaluer la qualité d'une caractéristique ou un sous-ensemble de caractéristiques en exploitant les informations intrinsèques des données d'entraînement sans impliquer un classificateur. Certains critères indépendants populaires sont des mesures de distance, des mesures d'information, des mesures de dépendance, et des mesures de cohérence (Liu and Motoda, 2012; Almuallim and Dietterich, 1994; Ben-Bassat, 1982; Hall, 2000).

Critères dépendants d'un classificateur Un critère dépendant est utilisé dans le modèle wrapper et qui nécessite un algorithme de classification prédéterminé dans le processus de sélection et utilise les performances de cet algorithme appliqué sur le sous-ensemble sélectionné. Ils donnent généralement des performances supérieures où il constate les caractéristiques les mieux adaptées à l'algorithme de classification prédéterminé, mais il a aussi tendance à être plus coûteux en calcul, et peut ne pas convenir à d'autres classificateurs (Blum and Langley, 1997). Par exemple, dans une tâche de classification, la précision de prédiction est largement utilisée comme la principale mesure et ainsi peut être utilisée en tant que critère d'évaluation dépendant pour la sélection de caractéristiques.

Critère d'arrêt

Un critère d'arrêt détermine le moment où le processus de sélection de caractéristiques devrait cesser. Certains critères d'arrêt fréquemment utilisés sont les suivants:

- La recherche est terminée.
- Un nombre minimum de caractéristiques ou le nombre maximum d'itérations est atteint.
- L'addition ultérieure (ou la suppression) de toute caractéristique ne produit pas un meilleur sous-ensemble (cas de stagnation).
- Un suffisamment bon sous-ensemble est sélectionné (par exemple, un sous-ensemble peut-être suffisamment bon si son taux d'erreur de classification est inférieur au taux d'erreur admissible pour une tâche donnée).

Validation

Un moyen simple pour la validation du résultat est de mesurer directement le résultat en utilisant la connaissance a priori sur les données. Si nous connaissons les caractéristiques pertinentes à l'avance, comme dans le cas de données de synthèse, on peut comparer cet ensemble connu de caractéristiques avec celles sélectionnées. Les connaissances sur les caractéristiques non pertinentes ou redondantes peuvent aussi aider. Nous ne nous attendons pas à ce qu'elles soient sélectionnées. Cependant, dans les applications du monde réel, nous ne disposons généralement pas d'une telle connaissance préalable. Par conséquent, nous devons compter sur des méthodes indirectes en suivant l'évolution des performances de classification et de prédiction avec le changement de caractéristiques. Par exemple, si nous utilisons le taux d'erreur de classification en tant qu'indicateur des performances pour une tâche de classification, pour un sous-ensemble de caractéristiques sélectionné. Nous pouvons tout simplement mesurer et comparer le taux d'erreur du classificateur entrainé sur le jeu de données complet avec celui entrainé sur un jeu de données qui inclut seulement le sous-ensemble de caractéristiques sélectionnées (Liu and Motoda, 2012; Witten and Frank, 2005).

3.3.3 Approches de sélection de caractéristiques

Il existe différentes méthodes de sélection de caractéristiques. En fonction de la façon dont ils combinent la recherche du sous ensemble à sélectionner avec la construction d'un modèle de classification (classificateur), nous pouvons principalement les divisées en cinq catégories (Saeys *et al.*, 2007; George and Raj, 2011; Li *et al.*, 2004), décrites ci-dessous. Nous mettons l'accent dans cette taxonomie sur les méthodes proposées pour l'identification de biomarqueurs à partir des données d'expression génique (ou chaque gène représente une caractéristique et chaque échantillon représente un exemple d'apprentissage).

Méthodes filtres (Filter methods)

Les méthodes filtres évaluent le pouvoir discriminant de gènes basées uniquement sur les propriétés intrinsèques de données d'expressions. De manière générale, ces méthodes estiment un score de pertinence, par la suite un système de seuillage est utilisé pour sélectionner les gènes les plus pertinents (Guyon and Elisseeff, 2003). Ce groupe de techniques est indépendant de tout système de classification, mais dans des conditions particulières, ils pourraient fournir l'ensemble optimal de caractéristiques pour un classifi-

cateur donné (Inza *et al.*, 2004). Les méthodes de cette catégorie peuvent être divisées en deux sous classes, citées ci-dessous.

Méthodes univariées (Univariate) Ici chaque gène est considéré séparément, ignorant ainsi les dépendances entre les gènes. Ce qui peut conduire à une mauvaise classification par rapport à d'autres types de techniques de sélection de caractéristiques. Les méthodes univariées sont divisées en deux sous-types:

- **Méthodes paramétriques** représente l'ensemble de techniques qui sont paramétrique (Jafari and Azuaje, 2006; Baldi and Long, 2001).
- **Méthodes non-paramétriques (Model-free)** représente l'ensemble de techniques qui ne sont pas paramétrique, mais capables de capturer des gènes nettement dérégulés dans un sous-ensemble d'échantillons (Thomas *et al.*, 2001; Breitling *et al.*, 2004). Ces méthodes offrent une approche plus spécifique pour l'identification de biomarqueurs et peuvent sélectionner des gènes présentant des motifs complexes.

Avantages

- Rapide,
- Evolutive,
- Indépendant d'un classificateur.

Inconvénients

- Ignore les dépendances (la corrélation) entre les gènes, ce qui réduit, dans certains cas, l'utilité des gènes sélectionnés pour la classification.
- Ignore l'interaction avec le classificateur.

Exemples

- En 1982, X2, « Information gain », « Gain ratio » (Ben-Bassat, 1982),
- En 2001, « Gamma » (Newton *et al.*, 2001),
- En 2001, méthodes de régression, « Wilcoxon rank sum » (non paramétrique) (Thomas *et al.*, 2001),
- En 2006, « Euclidean distance », « i-test » (Hu *et al.*, 2006),

- En 2006, « Bayesian methods » (Baldi and Long, 2001; Fox and Dimmic, 2006),
- En 2006, « t-test » et « ANOVA » (Jafari and Azuaje, 2006),
- En 2012, «Binary Matrix Shuffling Filter» (BMSF)(Zhang *et al.*, 2012),
- En 2013, méthode de sélection Basée sur un seuil (Van Hulse *et al.*, 2012),
- En 2013, RT-PLSDA qui est une combinaison entre « Partial least squares discriminant analysis » et « Randomization test » (Mao *et al.*, 2013),
- En 2013, « dynamic weighting-based feature selection algorithm» (DWFS) (Sun *et al.*, 2013).

Méthodes multivariées (Multivariate) En vue de surmonter le problème des dépendances entre les gènes soulevé dans les méthodes univariées, un certain nombre de techniques de filtrage multivariées ont été mises en place, visant l'intégration des dépendances entre les gènes à un certain degré (Hall, 1999).

Avantages

- Meilleurs complexité de calcul que les méthodes d'encapsulation (wrappers),
- Indépendant du classificateur.

Inconvénients

- Plus lent que les techniques univariées,
- Moins évolutif que les techniques univariées,
- Ignore l'interaction avec le classificateur.

Exemples

- En 2003, "Minimum Redundancy-Maximum Relevance" (MRMR) (Ding and Peng, 2005),
- En 2003, "Uncorrelated Shrunken Centroid" (USC) (Yeung *et al.*, 2003),

- En 2003, "Markov blanket" (Gevaert *et al.*, 2006; Xing *et al.*, 2001),
- En 2004, "Fast correlation-based feature selection" (FCBF) (Yu and Liu, 2004),
- En 2005, "Correlation-based feature selection"(CFS) (Wang *et al.*, 2005),
- En 2010, "Feature Selection Using Entropy Measure » (Zhu *et al.*, 2010a).

Méthodes enveloppes (Wrapper methods)

Les méthodes enveloppes sélectionnent le sous-ensemble de caractéristiques le plus discriminant en réduisant au minimum l'erreur de prédiction d'un classificateur particulier. Ces méthodes sont dépendantes d'un classificateur et elles sont principalement critiquées en raison de leurs énormes besoins en temps de calcul (Zhang *et al.*, 2012; Kohavi and John, 1997). Plus que cela, il n'y a aucune garantie que la solution proposée sera optimale si un autre classificateur est utilisé pour la prédiction. Selon la technique de recherche utilisée les méthodes « Wrapper » peuvent être divisées en deux catégories, citées ci-dessous.

Méthodes déterministes Ici l'algorithme de génération de sous-ensembles utilisé est déterministe.

Avantages

- Simple,
- Interagit avec le classificateur,
- Les modèles comportent des dépendances,
- Moins de calcul intensif que les méthodes aléatoires.

Inconvénients

- Risque le sur-apprentissage (overfitting), lorsque le nombre d'échantillons est petit.
- Plus enclins que les algorithmes randomisés pour se retrouver dans un optimum local (recherche gloutonne),
- La sélection dépend du classificateur.

Exemples

- En 1978, "Sequential forward selection" (SFS) (Kittler *et al.*, 1978),
- En 1978, "Sequential backward elimination" (SBE) (Kittler *et al.*, 1978),
- En 1994, "Plus q take-away r" (Ferri *et al.*, 1994),
- En 1998, "Beam search" (Egmont-Petersen *et al.*, 1998),
- En 2004, "Sequential search" (Inza *et al.*, 2004).

Méthodes aléatoires Ici l'algorithme de génération des sous-ensembles utilisé est stochastique.

Avantages

- Moins sujet à des optima locaux,
- Interagit avec le classificateur,
- Les modèles comportent des dépendances.

Inconvénients

- Plus de Risque de sur-apprentissage que les algorithmes déterministes,
- La sélection dépend du classificateur,
- Forte intensité de calcul (Inza *et al.*, 2004; Xing *et al.*, 2001).

Exemples

- En 2004, "Estimation of distribution algorithms" (Blanco *et al.*, 2004).
- En 2005, "Genetic algorithms " (Jirapech-Umpai and Aitken, 2005; Li *et al.*, 2001),
- En 2005, " Population-based metaheuristic", " randomized search heuristics " (Jirapech-Umpai and Aitken, 2005),
- En 2012, "Multi-Objective Particle Swarm Optimisation" (PSOFS) (Xue *et al.*, 2012).

Méthodes intégrées (Embedded methods)

Les méthodes intégrées représentent une autre classe de techniques, dans lesquelles la sélection de gènes est effectuée dans le processus d'entrainement au moment de la construction du model de classification. Ils sont généralement spécifiques à la technique de classification utilisée. Les méthodes Intégrées permettent encore des interactions avec l'algorithme d'apprentissage, mais le temps de calcul est plus petit que les méthodes d'encapsulation (Quinlan, 1986).

Avantages

- Interagit avec le classificateur,
- Meilleure en termes de complexité de calcul que les méthodes d'encapsulation (wrappers),
- Les modèles comportent des dépendances.

Inconvénients

- La sélection dépend du classificateur.

Exemples

- En 2002, "Feature selection using the weight vector of SVM" (Guyon *et al.*, 2002),
- En 2005, "Weights of logistic regression" (Ma and Huang, 2005),
- En 2006, "Random forest" (Díaz-Uriarte and De Andres, 2006),
- En 2007, SVM-RFE (Tang *et al.*, 2007).
- En 2011, SVM-RFE a été améliorée avec un "RBF kernel", qui est basée sur l'élimination récursive d'entité (SVM-RBF-RFE) (Liu *et al.*, 2011).

Méthodes d'ensemble (Ensemble methods)

Les méthodes d'ensemble représentent une classe relativement nouvelle des méthodes de sélection de caractéristique. Elles ont été proposées pour faire face aux problèmes d'instabilité observés dans de nombreuses techniques de sélection, lorsque de petites perturbations dans l'ensemble d'entrainement se

produisent. Ces méthodes sont basées sur des différentes stratégies de sous-échantillonnage, et un procédé simple qui exécute sur un certain nombre de sous-échantillons l'algorithme de sélection et les gènes obtenus sont fusionnés en un sous-ensemble plus stable (Yang *et al.*, 2010; Haury *et al.*, 2011; Kotsiantis, 2011).

Avantages

- Combine les efforts de plusieurs techniques de sélection de caractéristiques et de classificateur.
- Sélectionne des sous-ensembles de gènes plus stables.

Inconvénients

- La sélection dépend du classificateur,
- Forte intensité de calcul.

Exemples

- En 2008, "Robust feature selection using ensemble feature selection techniques" (Saeys *et al.*, 2008),
- En 2009, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods" (Abeel *et al.*, 2010),
- En 2010, "Ensemble gene selection by grouping for microarray data classification" (Liu *et al.*, 2010),
- En 2012, "Feature Selection-based Ensemble Method" (Namsrai *et al.*, 2013).

Méthodes hybrides (Hybrid methods)

Les méthodes hybrides utilisent généralement l'information de classement obtenue par les méthodes filtres, pour guider la recherche dans les algorithmes d'optimisation utilisées dans les méthodes wrapper. En plus, c'est une approche récente et une direction prometteuse dans le domaine de la sélection de caractéristiques (Liu and Zhou, 2003; Liu *et al.*, 2005; Ng and Chan, 2005). Dans le modèle hybride, un filtre est d'abord utilisé pour présélectionner une majorité de gènes (pertinents) de l'ensemble originale, ce qui donne un sous-ensemble filtré d'une taille relativement petite. Ensuite,

une méthode wrapper est appliquée pour sélectionner des gènes à partir de ce sous-ensemble filtré, afin d'optimiser la précision de prédiction du classificateur. Etant donné que le filtre réduit efficacement la taille de l'ensemble de gènes, la complexité du calcul ultérieur dans la méthode wrapper devient acceptable.

Avantages

- La simplicité des approches filtres joue un rôle primordial pour le dépistage de l'ensemble de gènes initial.
- L'utilisation de l'approche « wrapper » pour optimiser la précision de la classification du sous ensemble de gènes final sélectionné.

Inconvénients

- La sélection dépend du classificateur,
- Forte intensité de calcul.

Exemples

- En 2010, "Multiple-Filter-Multiple-Wrapper" (MFMW) (Leung and Hung, 2010),
- En 2010, "Fuzzy Random Forest Features selection" FRF-fs (Cadenas *et al.*, 2013),
- En 2013, "Greedy Randomized Adaptive Search Procedure" (GRASP) (Bermejo *et al.*, 2011).

3.4 Modèle d'îles généralisé

Le développement des capacités de calcul parallèle sous la forme de plusieurs CPUs et GPUs a motivé les chercheurs pour réfléchir à la façon d'utiliser ces capacitées omniprésentes pour faire face à des problèmes complexes qui exigent beaucoup de temps de calcul. Dans le cadre de l'optimisation, le modèle d'iles généralisé (GIM) (Izzo *et al.*, 2012) a été proposé comme une extension du modèle d'iles (IM) qui est basé sur l'algorithme génétique traditionnel (GA). Il a été démontré que le modèle d'iles pour l'optimisation coopérative et parallèle performe beaucoup plus que les approches globales avec la même quantité d'effort de calcul (Cantu-Paz, 2000). Pour les deux

modèles « GIM » et « IM », l'exécution en parallèle est effectuée non seulement pour accélérer le traitement, mais aussi pour exploiter les avantages de la coopération entre plusieurs métaheuristiques afin d'atteindre une meilleure convergence de l'algorithme globale d'optimisation.

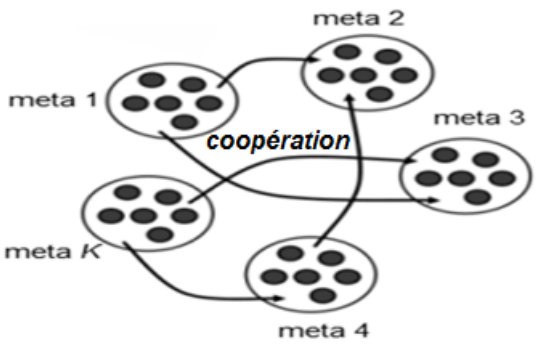


Figure 3.7: Modèle d'iles généralisé

Izzo et al. ont montré que le IM est un paradigme général qui peut être appliqué non seulement à des algorithmes génétiques ou évolutifs, mais aussi à une famille beaucoup plus large des algorithmes d'optimisation et des métaheuristiques, et même peut former des îles hétérogènes qui utilisent des algorithmes différents (comme représenté sur la Figure 3.7) (Izzo *et al.*, 2012). Les auteurs ont proposé un cadre général pour mettre en œuvre la coopération parallèle de méta-heuristiques et qui peut être appliqué à une large classe de problèmes d'optimisation. Le cadre proposé comprend un certain nombre de paramètres qui sont la mise en œuvre de la migration (synchrone ou asynchrone), nombre d'îles, la topologie de migration, l'intervalle de la migration, la politique de sélection de migration et la politique de remplacement de migration (Izzo *et al.*, 2012).

Partie II

Contributions

Chapitre 4

Approche de selection de caractéristiques inspirée du meta-learning et basée sur un ensemble de filtres pour la découverte de biomarqueurs du cancer ($MPME - FS$)

Contenu du chapitre

4.1 Introduction

Un défi majeur dans l'analyse des données d'expression génique est dû à leurs tailles: un très petit nombre d'échantillons, de l'ordre de plusieurs dizaines, contre des milliers de gènes associés à tous les échantillons. Ceci est communément connu comme "the curse-of-dimensionality" qui se caractérise également par un grand nombre de gènes redondants, bruyants et non pertinents ce qui influence l'efficacité du diagnostic (Bolón-Canedo *et al.*, 2014b). Les statistiques et les techniques d'apprentissage automatique ont été large-

ment utilisées pour l'identification des biomarqueurs, notamment la sélection de caractéristiques où les chercheurs essayent d'identifier les gènes les plus distinctifs qui peuvent atteindre de meilleure performance de prédiction des sous-types de cancer.
Un autre défi concerne les variations biologiques dans les tests cliniques réels qui nécessitent le développement des méthodes de sélection de caractéristiques d'une plus grande stabilité (He and Yu, 2010). La robustesse de la signature sélectionnée reste un objectif crucial dans la médecine personnalisée. Dans la biologie du cancer, il est très souhaitable d'utiliser une technique de sélection de caractéristique stable, ce qui peut réduire l'influence de ces variations biologiques dans les tests cliniques réels sur les patients.
Ainsi, une meilleure méthode de sélection pour la découverte de biomarqueurs doit tenir compte de deux aspects majeurs, la précision et la robustesse. Cette dernière peut être assurée par l'utilisation d'une méthode de sélection basée ensemble, contrairement à la précision qui peut être garantie en incorporant un modèle d'apprentissage dans le processus de sélection, ce qui représente les méthodes wrappers.

Cependant, la stabilité ou la robustesse des biomarqueurs est un aspect essentiel qui doit être pris en considération dans les méthodes de sélection de caractéristiques. A cet effet, il existe deux catégories principales de méthodes qui permettent une sélection plus robuste sans affecter la précision de prédiction dans le diagnostic du cancer (Khoshgoftaar *et al.*, 2013; Awada *et al.*, 2012). La première direction représente les méthodes de sélection de caractéristiques basées ensemble ("ensemble feature selection") (Abeel *et al.*, 2010; Yu *et al.*, 2012). L'autre direction représente les approches de sélection basées groupe ("group feature selection") (Liu *et al.*, 2010).
L'apprentissage basée ensemble est une technique robuste et populaire, en raison de l'immense succès de nombreuses méthodes d'ensemble dans les applications de bioinformatique. Il a l'avantage de surmonter le problème de dimensionnalité des données d'expression génique. Ainsi, il donne une plus grande précision et stabilité aux méthodes de sélection de caractéristiques que les techniques classiques ne peuvent pas atteindre. Par conséquent, l'utilisation de méthodes d'ensemble dans le problème de sélection de caractéristiques a été l'une des dernières tendances croissantes. Elles consistent à lancer plusieurs sélecteurs (le même ou différents) sur différents sous-échantillons, puis dans une deuxième étape agréger les résultats en utilisant une fonction de consensus afin de sélectionner le meilleur sous-ensemble final de biomarqueurs (Guan *et al.*, 2014). Un autre avantage de l'application des méthodes de sélection basées ensemble, qu'ils sont naturellement suscepti-

bles au parallélisme, ainsi que nous pouvons facilement entreprendre leurs paramètres en parallèle. La mise en œuvre parallèle de méthodes d'ensemble peut certainement accélérer le temps de calcul de sélection et permettre la résolution des problèmes à grande échelle en impliquant des multiprocesseurs pour l'exécution des différentes parties de l'ensemble en parallèle (Upadhyaya, 2013).

Récemment, un nouvel aspect dans l'apprentissage automatique a été proposé pour améliorer l'efficacité des classificateurs. Ce dernier est une extension des méthodes d'ensemble, nommé méta-apprentissage ou un ensemble d'ensembles. Le principe sous-jacent à cette technique est que les gènes sélectionnés ou bien classés par différents ensembles sont certainement plus pertinents que ceux sélectionnés par un seul ensemble (Yang *et al.*, 2010). Par conséquent, la bonne exploitation de ce concept par le bon choix de la fonction de consensus appropriée permet d'avoir des méthodes de sélection de caractéristiques plus efficaces et stables qui seraient concurrentiel et même surpassent celles existantes.
Dans cette direction, nous proposons une nouvelle méthode basée sur la notion du "meta-learning" pour la découverte de biomarqueurs les plus discriminants pour différents sous types de cancer. La méthode proposée, nommée "Massively parallel Meta-Ensemble Feature selection ($MPME-FS$)", peut être efficacement appliquée sur n'importe quel jeux de données ou en utilisant n'importe quelle méthode de classement (filtre). L'idée principale de notre approche est l'utilisation d'un ensemble d'ensembles (bagging) et la proposition de deux fonctions de consensus robustes afin de sélectionner des biomarqueurs à partir des données de puces à ADN. En outre, nous employons un ensemble de différents classificateurs à savoir les machines à vecteurs de support (SVM), K-plus proches voisins (KNN) en utilisant la validation croisée ("k-fold cross-validation") pour l'évaluation des sous-ensembles générés.

Dans un premier temps, ce chapitre met l'accent sur les différents aspects qui concernent l'application de méthodes des sélection de caractéristiques basées ensemble à la découverte de biomarqueurs à partir des données d'expression génique. Nous proposons ensuite une nouvelle méthode parallèle de sélection de caractéristiques basée sur la notion du meta-learning qui permet de sélectionner des biomarqueurs robustes et précis à partir d'ensembles de données de puces à ADN et qui peut être généralisée et appliquée sur plusieurs études génomiques. Deux types de filtre sont utilisés dans cette étude: "ReliefF" et "Information Gain". Nous discutons également les résultats en termes de robustesse, la puissance de classification et la signification

biologique des signatures sélectionnées.

4.2 L'application des méthodes de sélection de caractéristiques basées ensemble à la découverte de biomarqueurs

Par analogie avec les méthodes d'ensemble en apprentissage automatique supervisé qui combinent plusieurs classificateurs afin d'obtenir une grande précision de classification et de prédiction, telles que le bagging et le boosting (Yang *et al.*, 2010). La sélection de caractéristiques basée ensemble a également reçu beaucoup d'attention récemment.
Nous présentons principalement ici les différents aspects à considérer dans les méthodes de sélection de caractéristiques basées ensemble pour la découverte de biomarqueurs et qui peuvent aider les chercheurs à classer toute méthode d'entre elles.

Les principaux problèmes critiques dans cette catégorie de méthodes sont à la fois la construction de différents sélecteurs locaux et la fonction de consensus utilisée pour combiner les différents sous-ensembles de caractéristiques (He and Yu, 2010; Awada *et al.*, 2012). Par conséquent, le premier aspect à examiner est la conception de la diversité au sein de l'ensemble. Ce critère divise les méthodes de sélection de caractéristiques d'ensemble en trois classes:

- **Ensemble fondé sur la diversité de données (data diversity)** où le même sélecteur est lancé sur différents sous-échantillons générés à partir de l'ensemble de données d'origine (Saeys *et al.*, 2008; Abeel *et al.*, 2010).
- **Ensemble fondé sur la diversité fonctionnelle (functional diversity)** où différents sélecteurs sont lancés sur le même ensemble de données (sans échantillonnage) (Bolón-Canedo *et al.*, 2014a).
- **Ensemble fondé sur la diversité de données et la diversité fonctionnelle (data and functional diversity)** ici la diversité de données et fonctionnelle sont combinés dans lequel différents algorithmes de sélection de caractéristiques sont effectués sur différents sous-échantillons.

Un autre aspect à prendre en compte dans les méthodes de sélection de caractéristique d'ensemble est la représentation utilisée par les différents sélecteurs, étant donné que la notation des résultats n'est pas la même dans

tous les algorithmes de sélection de caractéristiques. Cela a un grand impact sur la fonction de consensus utilisée pour agréger les résultats (Saeys *et al.*, 2008). Typiquement, nous pouvons observer trois types de représentations:

- **Représentation en sous-ensemble de caractéristiques** l'information est transformée sous forme de sous-ensembles contenant les caractéristiques sélectionnées uniquement (généralement avec des tailles différentes)
- **Représentation en classement de toutes les caractéristiques** l'information est transformée sous forme d'un vecteur qui contient toutes les caractéristiques classées (un seuil est nécessaire pour la sélection)
- **Représentation en pondération de caractéristiques** l'information est transformée sous forme d'un vecteur à deux dimensions caractéristique/poids qui peut facilement être converti en une représentation en classement.

Des études récentes ont mis l'accent sur les méthodes d'ensemble en utilisant les sélecteurs basés wrapper (Yang *et al.*, 2013; Xu *et al.*, 2013). Cela, nous amène à considérer la dépendance des sélecteurs à tout classificateur comme un aspect important dans les méthodes de sélection de caractéristiques d'ensemble. Ce critère influe la qualité des solutions au sein de l'ensemble et le coût du calcul global de la sélection. Ainsi, il divise les méthodes de sélection de caractéristiques d'ensemble en:

- **Ensembles basés "filter"** ils sont simples, rapides et indépendants de tout classificateur (Bolón-Canedo *et al.*, 2014a).
- **Ensembles basés "wrapper"** ils sont très couteux en calcul et ont le risque de sur-apprentissage, en raison de la grande dimensionnalité des données. Ainsi, ils comprennent des interactions entre la recherche de sous-ensemble et caractéristiques et un classificateur (modèle d'apprentissage). En outre, ils ont la capacité de prendre en compte les dépendances entre les caractéristiques, ce qui est important dans quelques applications bioinformatique (Ghorai *et al.*, 2011).
- **Ensembles basés "embedded"** ils utilisent l'information interne du classificateur pour effectuer la sélection et montrent une meilleure complexité de calcul que les méthodes basées wrapper (Saeys *et al.*, 2008).
- **Ensembles hybrides** ils sont une combinaison de méthodes filter et wrapper qui utilisent l'information de classement obtenue en utilisant des filtres pour guider la recherche dans les algorithmes d'optimisation utilisés par la méthode wrapper (Xu *et al.*, 2013).

4.3 Principe de MPME-FS

4.3.1 Formulation du problème

Par analogie avec les modèles "meta-ensemble" dans l'apprentissage automatique supervisé (Yang *et al.*, 2010), l'approche proposée est conçue comme un ensemble d'ensembles de différents sélecteurs qui effectuent la sélection en parallèle à travers divers ensembles et deux fonctions de consensus (voir Figure 4.1). Dans ce qui suit, nous formulons d'abord le problème ainsi que la représentation de notre solution. Puis, le cadre général est exploré y compris la construction parallèle des listes de classement ainsi que les fonctions de consensus.

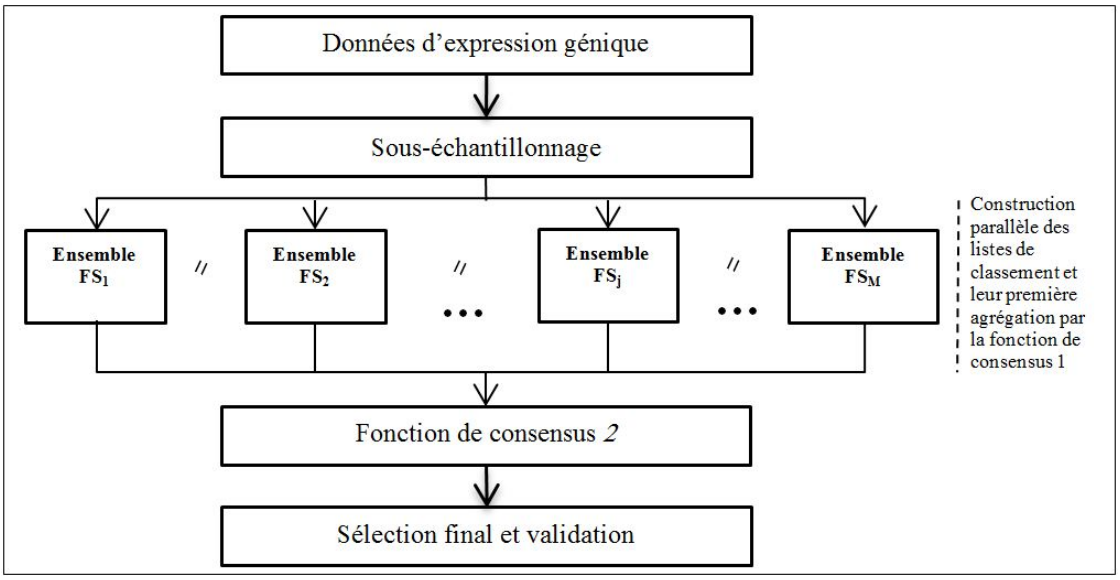


Figure 4.1: Modèle parallèle de $MPME - FS$

La découverte de biomarqueurs à partir de l'expression de gènes est le problème de la sélection d'un sous-ensemble de biomarqueurs les plus représentatifs à partir d'un grand ensemble de données. Étant donné un ensemble X de K caractéristiques, avec K très grand, le problème consiste à trouver le sous-ensemble minimal $X_s^* \subset X$ qui contient les caractéristiques les plus pertinentes et non redondantes. La sélection de caractéristiques basée ensemble est une technique prometteuse pour faire face aux structures complexes de données et permet d'atténuer les problèmes avec une petite taille d'échantillons et de dimension élevée.

L'utilisation d'un ensemble d'ensembles de filtres conduit à plusieurs sous-ensembles de biomarqueurs. Notons X_{sj}^{i} le sous-ensemble de caractéristiques j sélectionnés en utilisant le filtre i. Par conséquent, deux questions doivent être abordées. La première est liée à l'importance de chaque caractéristique et la seconde est liée à la manière dont les sous-ensembles sont agrégés pour arriver à sélectionner le sous-ensemble final de caractéristiques.

Afin de traiter correctement ces deux questions, nous proposons une approche en deux étapes qui utilise des filtres de classement et des fonctions de consensus. Lors d'une première étape plusieurs sous-ensembles de listes de classement des caractéristiques sont construits, en utilisant plusieurs filtres, puis l'agrégation de ces sous-ensembles est effectuée à deux niveaux pour former d'abord les ensembles puis le méta-ensemble. Plus formellement, la sortie de la première étape peut être représentée comme suit:

$$X_{sj}^{i} = \left\{(f_{ij}^{k}, w_{ij}^{k}) \quad ou \quad i,j = 1...(N,M) \quad et \quad k = 1...K\right\} \tag{4.1}$$

f_{ij}^{k} représente le rang de la caractéristique k dans l'ensemble j en utilisant un filtre i. Sa pertinence est donnée par le poids w_{ij}^{k}. Le poids global de la caractéristique k au sein de l'ensemble j est désigné par w_{j}^{k}.
Trois paires de sous-ensembles de caractéristiques et leurs poids correspondants sont nécessaires.

- Le premier est $Lbest_j$ qui représente les meilleures caractéristiques locales dans chaque ensemble j.

  $$Lbest_j = \left\{(k, w_{j}^{k}) \quad ou \quad j = 1...M \quad et \quad k = 1...K\right\} \tag{4.2}$$

  Ce sous-ensemble est le résultat d'un processus d'agrégation sur les sous-ensembles $X_{sj}^{i=1..N}$

- Le second est $Gbest$. Il représente les meilleures caractéristiques globales sélectionner au niveau du méta-ensemble. Il représente le résultat d'un processus d'agrégation sur les sous-ensembles $Lbest_j$.

- Le troisième est $Fbest = X_{s}^{*} \subset X$. Il représente les caractéristiques finales sélectionnées par la méthode proposée $MPME - FS$.

4.3.2 Le cadre général de $MPME - FS$

Le cadre général de la méthode proposée "$MPME - FS$" se compose de plusieurs ensembles de filtres qui s'exécutent en parallèle chacun d'entre

eux emploie une fonction de consensus robuste pour sélectionner les meilleurs biomarqueurs au sein de chaque ensemble. La prochaine étape est l'agrégation des résultats de tous les ensembles en utilisant une deuxième fonction de consensus. Enfin, sélectionner les biomarqueurs qui ont les scores les plus élevés, attribués par tous les filtres de tous les ensembles comme le montre la Figure 4.1.

Le processus de sélection commence par la construction de M sous-échantillons $S_{i=1...M}$ à partir de l'ensemble des données initial. Ensuite, la sélection parallèle est lancée dans tous les ensembles. A ce stade, chaque ensemble j construit N listes de biomarqueurs classés X^i_{sj}, en utilisant le $filtre_i$. Afin d'assurer l'objectif de la diversité fonctionnel et de données lors de la construction des listes X^i_{sj}, nous utilisons le partitionnement des données avec un chevauchement permettant la création d'un ensemble de données réduit pour chaque $filter_i$. La perturbation des données implique la génération de plusieurs sous-échantillons en éliminant certains échantillons selon un taux de chevauchement à partir de l'ensemble de données d'origine de manière aléatoire. Sachant que, le chevauchement (the overlap) représente le pourcentage d'échantillons appartenant à l'ensemble de données d'origine (Awada *et al.*, 2012).
Par la suite, une fonction de consensus au sein de chaque ensemble est appliquée afin d'agréger ces listes classées X^i_{sj} et finalement obtenir les meilleures biomarqueurs locaux dans l'ensemble j ($Lbest_j$), comme illustré dans la Figure 4.2. Il faut noter que la construction des listes de classement au sein de chaque ensemble est effectuée en parallèle et à travers différents filtres (Informations Gain, le ratio Gain, Ratio Fisher, incertitude symétrique, ReliefF). L'étape suivante consiste à l'agrégation des meilleures biomarqueurs locaux sélectionnés dans l'étape précédente à travers le méta-ensemble. Ensuite, construire le meilleur sous-ensemble global de biomarqueurs ($Gbest$), à l'aide de la fonction de consensus 2 avec l'accumulation des scores associés à ces gènes. Enfin, nous considérons les gènes les mieux classés qui constituent le sous-ensemble final $Fbest$.

4.3.3 Les fonctions de consensus

Le point clé dans les méthodes d'ensemble est la manière de combiner différents sous-ensembles conçus par plusieurs sélecteurs qui conduit à l'ensemble final de biomarqueurs les plus discriminants. Ce problème a reçu une attention considérable au cours des dernières années (Boulesteix and Slawski, 2009). Les méthodes d'agrégation dépendent de la représentation des sous-ensembles de sortis des sélecteurs, déjà discuté dans la section précédente,

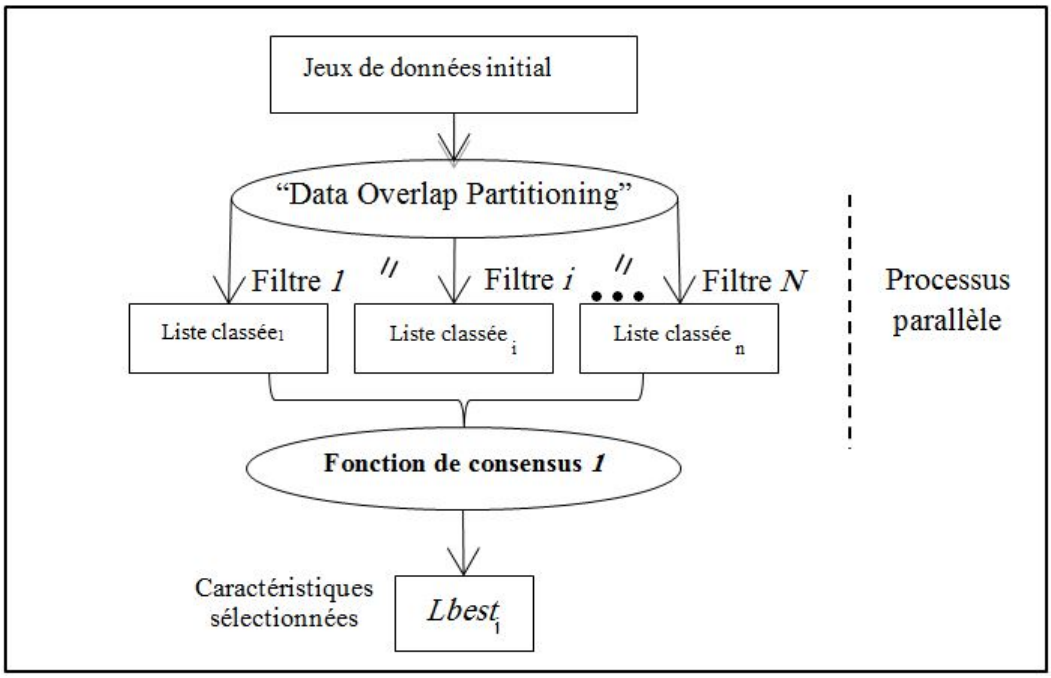


Figure 4.2: Sélection de caractéristiques à base d'ensemble

qui peut être divisée en trois types: sous-ensemble de caractéristiques ("feature subset"), classement de toutes les caractéristiques ("feature ranking") et pondération de caractéristiques ("feature weighting-score") (Awada *et al.*, 2012). Sur la base de ces représentations, il existe de nombreuses fonctions de consensus, à savoir le vote pondéré ("weighted voting"), l'agrégation basée sur un seuil ("threshold based aggregation") pour les deux représentations en classement et en pondération de caractéristiques et enfin le comptage des caractéristiques les plus fréquemment choisies pour la représentation en sous-ensemble (Saeys *et al.*, 2008; Haury *et al.*, 2011).

Certes, le choix de la fonction de consensus appropriée est une tâche difficile dans les méthodes d'ensemble. Dans notre travail, nous proposons deux fonctions de consensus; la première dans le niveau ensembles et la seconde fonction dans le niveau méta-ensemble. Les deux sont basées sur le classement de caractéristiques et leurs pondérations globales, ce qui conduit à une sélection finale plus robuste et plus parcimonieuse.

La première fonction de consensus permet l'agrégation des listes de classements générées par des filtres dans le même ensemble. Cette fonction est inspirée à la fois du principe des fonctions qui comptent les caractéristiques

les plus fréquemment choisies et les fonctions d'agrégation à base de vote pondéré, mais avec une sélection très stricte en appliquant l'intersection de l'ensemble des listes de classement et non pas un nombre fixe de fréquences. Aux fins d'intersection, nous utilisons un seuil désigné par $TS1$ afin de sélectionner seulement des gènes appartenant au $TS1$ premier classés. Ensuite, les poids de tous les gènes sélectionnés dans tout l'ensemble j désignés par W_{ij}^k sont accumulés pour obtenir les poids globaux de gènes sélectionnés W_j^k à travers tous les ensembles. Plus formellement, la fonction de $Consensus1$ peut être décrite comme suit:

$$\begin{cases} Lbest_j = \{(k, w_j^k)\} = \begin{cases} \bigcap_{i=1}^{N} \{(f_{i,j}^k, w_{i,j}^k)\} \\ et \quad f_{i,j}^k \leqslant TS1 \end{cases} \\ ou \quad w_j^k = \sum_{i=1}^{N} w_{i,j}^k \end{cases} \tag{4.3}$$

De cette façon, nous obtenons plusieurs ensembles $Lbest_j$ contenant les paires des meilleurs gènes sélectionnés avec leurs poids globaux dans chaque $ensemble_j$ $\{k, W_j^k\}$. Ce dernier sera l'entrée de la deuxième fonction de consensus dans le niveau méta-ensemble, afin de construire le sous-ensemble $Gbest$ qui contribue à la sélection finale.

La seconde fonction de consensus consiste principalement d'agrégation des M sous-ensembles $Lbest_j$ générés dans l'étape parallèle précédente. Le sous-ensemble $Gbest$ représente les paires de gènes et leurs poids accumulés appartenant à l'union des meilleurs sous-ensembles locaux M de biomarqeurs, qui peut être calculée comme suit:

$$\begin{cases} Gbest = \{(k, bw^k)\} = \bigcup_{j=1}^{M} Lbest_j \\ ou \quad bw^k = \sum_{j=1}^{M} w_j^k \end{cases} \tag{4.4}$$

Enfin, nous sélectionnons les gènes qui apartient au meilleurs premiers $TS2$ gènes classés à partir de l'ensemble $Gbest$ et qui représentent les gènes finaux sélectionnés pour être validés dans l'étape de validation. Le pseudo-code de l'ensemble du processus peut être résumé comme suit:

Algorithm 1 ***"Massively parallel Meta-Ensemble Feature Selection"*** $(MPME - FS)$

Parameters:
N: ensemble size
M: meta-ensemble size
$TS1$: threshold of intersection
$TS2$: percentage of best selected features in meta-ensemble
$Overlap$: overlap of data sampling
Input:
D: dataset with K features
L: sample labels in D
Output:
$Fbest$: final best selected features
Parameters initialization:
$Gbest, Lbest = \emptyset; Filter = filteri;$
Parallel Meta ensemble feature selection process:
1: For (each $ensemble_j$ / $j = 1...M$) do in parallel
// determining $Lbest_j$ for each $ensemble_j$
2: For ($i = 1...N$) do in parallel
3: S_i = sampling $(D, Overlap)$
4: $(k, bw^k) = Filter(S_i, L)$
5: End For in parallel
// obtaining the rank of each feature in the ensemblej
6: $\{(f_{i,j}^k, w_{i,j}^k)\}$ = Sort $(\{(k, bw_{i,j}^k)\})$
// aggregating results of all filters within the $ensemble_j$
7: $Lbest_j = \{(k, bw^k)\}$ = Consensus function 1 $(\{(f_{i,j}^k, w_{i,j}^k)\})$
8: End For in parallel
// aggregating all $Lbest_j$ subsets
9: $Gbest$ = Consensus function 2 $(Lbest_j)$
// final selection
10: Sort $(Gbest)$ // sort k based on bw^k
11: $Fbest$ = select $TS2$ best features from $Gbest$

4.4 Résulats et discussions

Dans cette section, l'analyse de performances de classification et de robustesse ainsi que l'interprétation biologique des résultats de $MPME - FS$ sont présentés. Tout d'abord, les ensembles de données et les paramètres expérimentaux utilisés dans cette analyse sont brièvement décrits. Deuxièmement, nous analysons les performances de classification en termes de précision (Accuracy), sensibilité (sensitivity) et la spécificité (specificity) en utilisant différents classificateurs. Après cela, nous étudions la robustesse des signatures sélectionnées. Enfin, nous effectuons une interprétation biologique des biomarqueurs sélectionnés.

Données	#Gènes	#Classes	#Echantillons
Ovarian	15.154	2	253
Leukemia	7.129	2	72
DLBCL	5.469	2	77
Colon	2.000	2	62
SRBCT	2.308	4	83

Table 4.1: Caractéristiques des différents ensembles de données utilisés

4.4.1 Paramètres et jeux de données (GED)

Toutes les expérimentations ont été effectuées en utilisant l'environnement Parallel Computing Toolbox de MATLAB® (PCT). La méthode proposée $MPME - FS$ a été évaluée au moyen de cinq ensembles de données accessibles au public des puces à ADN et qui peuvent être divisés en deux types binaires et multiclasses. Les ensembles de données binaires sont les plus importants et peuvent séparer les patients sains des patients atteints de cancer, tandis que les ensembles de données multiclasses sont utilisés pour différencier les différents types de cancers. Ainsi, les ensembles de données ont été récupérés à partir de deux dépôts de données: "Kent Ridge bio-medical data repository"[1] et "Gene Expression Model Selector", de "Vanderbilt University"[2]. Les caractéristiques principales des jeux de données sont présentées dans le Tableau 4.1.

[1] `http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html`

[2] `http://www.gems-system.org`

Pour évaluer les performances de notre méthode parallèle de sélection de caractéristiques basée méta-ensemble, nous utilisons dans nos expérimentations deux filtres largement connus et réussis: Information Gain et ReliefF. Sur la base d'une évaluation empirique en utilisant différents paramètres de la méthode proposée, le meilleur réglage des paramètres de $MPME-FS$ et qui est adopté dans cette étude est représenté dans le Tableau 4.2.

Filtres	InfoGain, ReliefF
Taille de l'ensemble	10
Taille du meta-ensemble	100
$TS1$	150
$TS2$	30
Overlap	80
Valeur du K dans ReliefF	10

Table 4.2: Réglage de paramètres du $MPME-FS$

4.4.2 Analyse de performances de classification et de prédiction

La première expérimentation est consacrée à lévaluation des performances de $MPME-FS$ en termes de précision, sensibilité, spécificité et le nombre de biomarqueurs sélectionnés à l'aide de la technique de validation croisée (k-fold cross validation, k =10). Cette dernière est un choix commun dans la littérature spécialisée (Zhu *et al.*, 2010b), qui divise l'ensemble des données en k sous-ensembles aléatoirement pour évaluer la précision de la signature sélectionnée à travers plusieurs itérations.

Afin d'assurer une bonne évaluation des capacités de classification des gènes sélectionnés, nous utilisons d'abord différents classificateurs séparément (SVM, KNN, ANN). Ensuite, nous utilisons un ensemble de différents classificateurs (SVM, KNN et ANN) avec un vote majoritaire comme fonction de consensus. Ainsi, les résultats indiqués dans la Figure 4.3 représentent les précisions moyennes de $MPME-FS$ données par SVM, KNN, ANN et l'ensemble de classificateurs décrits ci-dessus. Pour des raisons de comparaison, nous utilisons dans cette expérimentation deux filtres information Gain et ReliefF pour effectuer la sélection dans l'approche proposée. A partir de

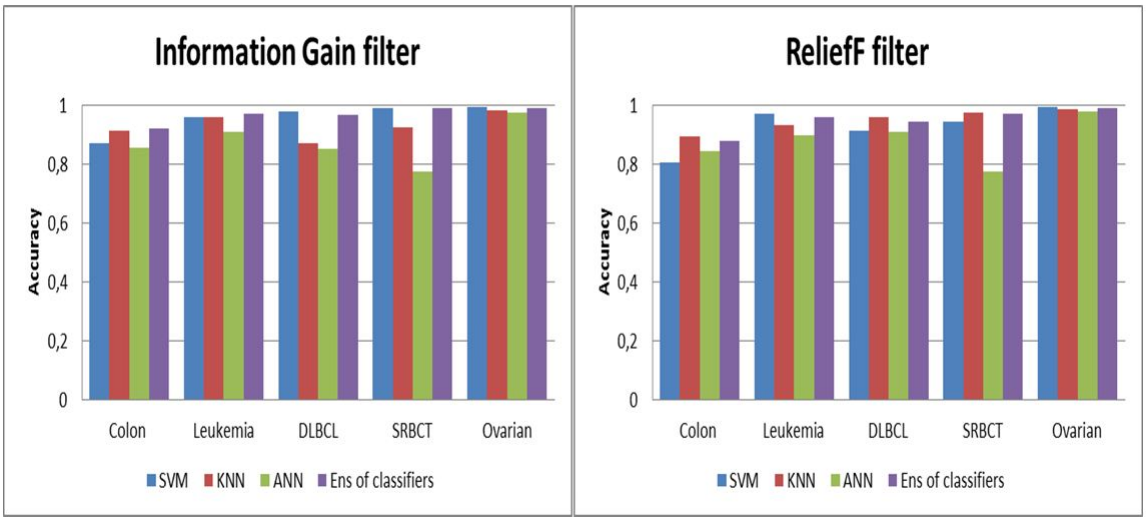


Figure 4.3: La Précision moyenne de classification (validation croisée, k=10) de $MPME-FS$ en utilisant: SVM, KNN, ANN et un ensemble de différents classificateurs à travers les cinq ensembles de données et pour les deux filters InfoGain et ReliefF

cette figure, nous observons que la méthode $MPME-FS$ fonctionne mieux en utilisant le filtre Information Gain que le filtre ReliefF à travers les cinq ensembles de données.

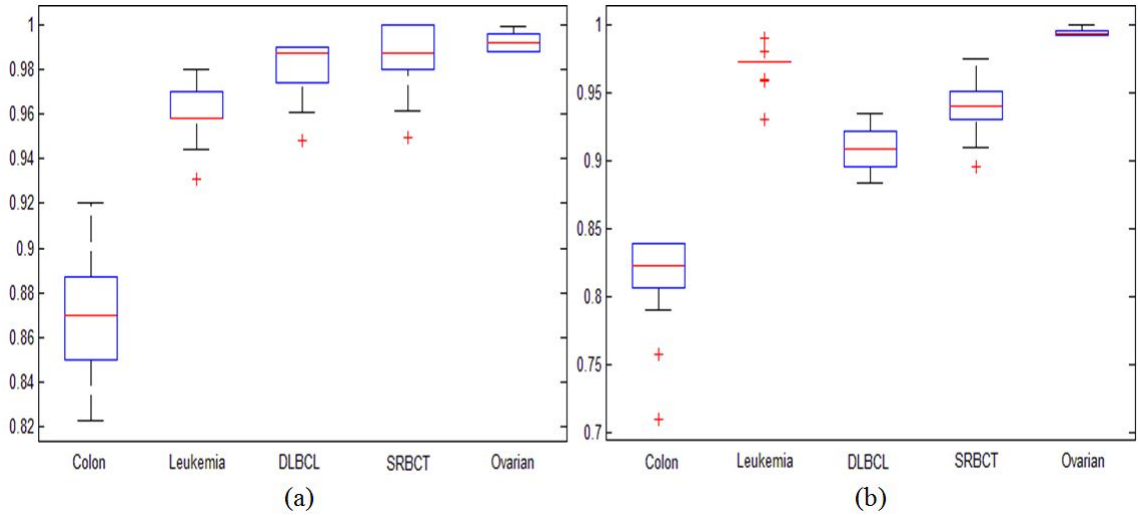


Figure 4.4: Boxplots de $MPME-FS$ sur les jeux de données colon, leukemia, DLBCL, SRBCT et ovarian à travers 30 exécutions. pour les deux filtres : (a) Information Gain et (b) ReliefF

Une deuxième observation qui peut être soulevée est que l'ensemble du classificateur donne une plus grande précision que les classificateurs simples, ce qui n'est pas surprenant car il combine les efforts des trois classificateurs.

En outre, la Figure 4.4 montre les boxplots des résultats de précision de $MPME-FS$ en utilisant le SVM comme classificateur et réalisés à travers trente exécutions pour les deux filtres Information Gain et ReliefF (Figure. 4.4.a, b successivement). La variance en utilisant les deux filtres entre les différentes exécutions est très faible (entre 0,001 et 0,04) ce qui indique que la méthode proposée est stable.

Nous présentons également dans le Tableau 4.3 les résultats moyens en termes de la sensibilité, la spécificité et le nombre de biomarqueurs sélectionnés de $MPME-FS$ en utilisant les différents classificateurs. Toutes les expériences précédentes ont été effectuées sur 30 exécutions indépendantes afin d'avoir des conclusions statistiquement significatives.

		SVM		**KNN**		**ANN**		**Ensemble de classificateur**		**# gènes**
		Sensi	Speci	Sensi	Speci	Sensi	Speci	Sensi	Speci	
InfoGain	**Colon**	0.875	0.804	0.925	0.907	0.871	0.831	0.914	0.909	30
	Leukemia	0.957	1	0.978	0.96	0.934	0.92	0.953	1	31
	DLBCL	0.965	1	0.827	1	0.924	0.63	0.952	1	27
	SRBCT	1	0.94	0.896	1	0.89	0.92	1	0.963	32
	Ovarian	0.998	1	0.993	0.967	1	0.978	1	0.988	39
Moyenne		**0,959**	**0,948**	**0,923**	**0,966**	**0,923**	**0,855**	**0,963**	**0,972**	**31**
ReliefF	**Colon**	0.875	0.81	0.925	0.863	0.9	0.818	0.89	0.863	31
	Leukemia	0.878	0.96	1	0.84	0.872	0.88	1	0.96	33
	DLBCL	0.931	0.947	0.965	0.947	0.948	0.842	0.948	0.947	27
	SRBCT	0.931	0.944	0.965	1	0.931	0.925	0.965	0.981	33
	Ovarian	1	1	1	0.978	1	0.978	1	0.989	41
Moyenne		**0,923**	**0,932**	**0,971**	**0,925**	**0,9302**	**0,888**	**0,960**	**0,948**	**33**

Table 4.3: Résultats de la classification moyennes en termes de sensibilité (Sensi), spécificité (spécifique) et le nombre de biomarqueurs sélectionnés (#genes) de MPME-FS utilisant à la fois les filtres: Information Gain et ReliefF à travers des cinq ensembles de données.

4.4.3 Analyse de robustesse

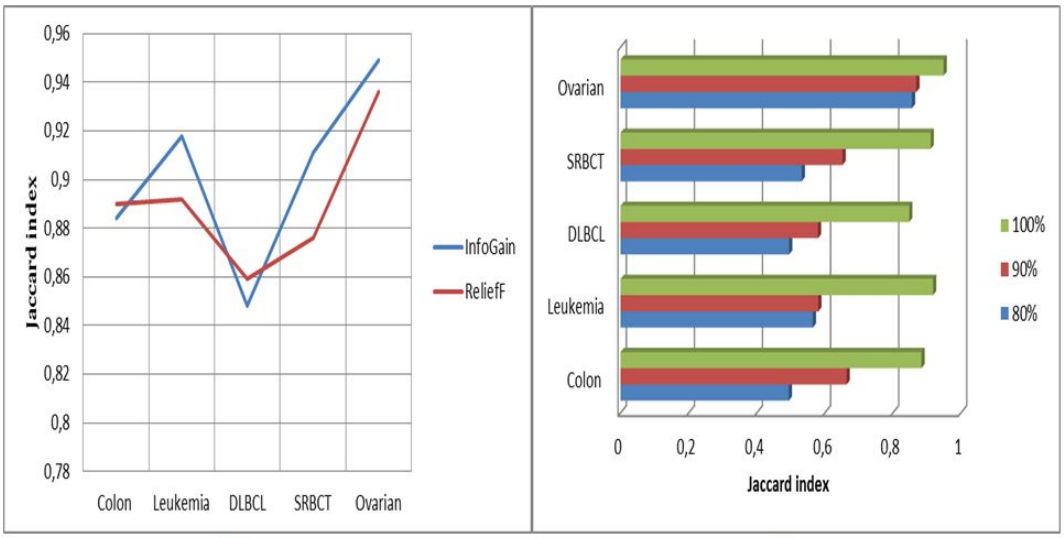
Nous explorons et discutons également dans cette étude la robustesse de la signature sélectionnée par l'approche $MPME-FS$. Donc, nous évaluons

la similitude entre les sorties des différentes exécutions indépendantes de notre méthode. La stabilité globale est définie comme étant la moyenne des similarités de toutes les paires de combinaisons possible entre les différents sélecteurs et définie comme suit (Saeys *et al.*, 2008):

$$S_{tot} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S(f_i, f_j)}{k(k-1)} \tag{4.5}$$

où f_i représente le résultat de la méthode de sélection de caractéristiques appliquée à un $sous - chantillon_i$ $(1 \leqslant i \leqslant 20)$, et $S(f_i, f_j)$ représente une mesure de similarité entre f_i et f_j. Pour une représentation de sortie en sous-ensembles (comme dans notre cas), nous utilisons l'indice de Jaccard (JI) comme mesure et qui peut être calculée comme suit:

$$S(f_i, f_j) = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} \tag{4.6}$$



(a) (b)

Figure 4.5: Résultats de robustesse de MPME-FS en terme de l'indice de Jaccard à travers 20 exécutions indépendants pour les cinq ensembles de données. (a) par rapport aux deux filtres Information Gain et ReliefF (b) par rapport aux taux de perturbations de sous-échantillonnage entre les différents éxécutions indépendants (80, 90 et 100%)

Cette partie d'expérimentation évalue la stabilité globale de la signature sélectionnée à travers les cinq ensembles de données à l'aide de deux filtres

Information Gain et ReliefF, qui est présentée dans la Figure 4.5.a. Les résultats de la stabilité globale en terme de Jaccard index indiquent que la méthode $MPME - FS$ réalisée en utilisant le filtre Information Gain est généralement plus robuste sur la plupart des ensembles de données. Afin de fournir une meilleure analyse de la robustesse, nous évaluons également l'effet du taux de perturbations des données lors de la création des sous-échantillons sur la stabilité globale des signatures. Dans cette expérience, nous utilisons le filtre Information Gain à travers les cinq ensembles de données dont les résultats peuvent être vus dans la Figure 4.5.b, qui indique que la robustesse diminue à chaque fois que le taux de perturbation est faible.

Indice de Gène	Numéro d'accession	Description de Gène
66	T71025	3' UTR 1 84103 Human (HUMAN).
1423	J02854	gene 1 "MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN); contains element TAR1 repetitive element.
1414	R64115	3' UTR 2a 139618 ADENOSYLHOMOCYSTEINASE (Homo sapiens).
137	D25217	gene1 "Human mRNA (KIAA0027) for ORF, partial cds.
138	M26697	gene 1 "Human nucleolar protein (B23) mRNA, complete cds.
241	M36981	gene 1 "Human putative NDP kinase (nm23-H2S) mRNA, complete cds.
245	M76378	gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6.
249	M63391	gene 1 "Human desmin gene, complete cds.
267	M76378	gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6.
1843	H06524	3' UTR 1 44386 "GELSOLIN PRECURSOR, PLASMA (HUMAN).
286	H64489	3' UTR 2a 238846 LEUKOCYTE ANTIGEN CD37 (Homo sapiens).
365	X14958	gene 1 Human hmgI mRNA for high mobility group protein Y.
377	Z50753	gene 1 H.sapiens mRNA for GCAP-II/uroguanylin precursor.
1960	D59253	gene 1 Human mRNA for NCBP interacting protein 1.
493	R87126	3' UTR 2a 197371 "MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
513	M22382	gene 1 MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN).
625	X12671	gene 1 Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1.
739	X12369	gene 1 "TROPOMYOSIN ALPHA CHAIN, SMOOTH MUSCLE (HUMAN).
897	H43887	3' UTR 2a 183264 COMPLEMENT FACTOR D PRECURSOR (Homo sapiens).
765	M76378	gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6.
780	H40095	3' UTR 1 175181 MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN).
812	Z49269	gene 1 H.sapiens gene for chemokine HCC-1.
964	T86473	3' UTR 1 114645 NUCLEOSIDE DIPHOSPHATE KINASE A (HUMAN).
1042	R36977	3' UTR 1 26045 P03001 TRANSCRIPTION FACTOR IIIA.
1411	H77597	3' UTR 1 214162 H.sapiens mRNA for metallothionein (HUMAN).
1494	X86693	gene 1 H.sapiens mRNA for hevin like protein.
1582	X63629	Gene 1 H.sapiens mRNA for p cadherin.
1635	M36634	Gene 1 "Human vasoactive intestinal peptide (VIP) mRNA, complete cds.
1771	J05032	Gene 1 "Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds.
1263	T40454	3' UTR 2a 60221 ANTIGENIC SURFACE DETERMINANT PROTEIN OA3 PRECURSOR (Homo sapiens)

Table 4.4: Description des trente tops premiers gènes sélectionnés pour le cancer du Colon, avec un nombre de fréquence complet (Freq = 30), à tavers 30 essais indépendants

4.4.4 L'interprétation biologique des biomarqueurs découverts

Nous abordons ici l'analyse biologique des biomarqueurs sélectionnés. Nous nous concentrons dans cette expérience sur l'analyse des biomarqueurs sélectionnés pour les Cancers du Colon et Leucémie qui sont largement étudiés dans la littérature. Ainsi, les Tableaux 4.4 et 4.5 listent et décrivent les trente tops gènes classés à travers plus de 30 essais indépendants et qui ont un niveau de fréquence complet (freq = 30) des deux types du Cancer, Colon et Leucémie, successivement. En outre, les gènes sélectionnés sont considérés comme informative par d'autres méthodes connues dans la littérature. Spécialement, dans le cas de Leucémie, qui a été largement étudié dans ce domaine.

En conséquence, les gènes sélectionnés pour Leucémie figurant en caractères gras dans le Tableau 4.5, ont été également sélectionnés parmi les 25 tops gènes les plus pertinentes par Wu et al. (Wu *et al.*, 2012).
En outre, huit de ces trente premiers gènes sélectionnés par notre méthode, à savoir, M23197_at, M27891_at, U05259_rna1_at, U46499_at, X95735_at, L09209_s_at, M31523_at et M89957_at ont été jugés comme pertinents par Zhu et al. (Zhu *et al.*, 2010b). Les gènes sélectionnés peuvent maintenant être validés par les biologistes à travers des essais cliniques. Nous prévoyons que ces découvertes peuvent offrir des informations utiles pour les biologistes et les experts médicaux.

Indice de Gène	Numéro d'accession	Description de Gène
758	**D88270_at**	GB DEF = (lambda) DNA for immunoglobin light chain
760	D88422_at	CYSTATIN A
1144	**J05243_at**	SPTAN1 Spectrin, alpha, nan-erythrocytic 1 (alpha-fodrin)
1630	L47738_at	Inducible protein mRNA
1685	**M11722_at**	Terminal transferase mRNA
1834	**M23197_at**	CD33 CD33 antigen (differentiation antigen)
1882	**M27891_at**	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
1902	M29474_at	Recombination activating protein (RAG-1) gene
2121	**M63138_at**	CTSD Cathepsin D (lysosomal aspartyl protease)
2128	M63379_at	CLU Clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)
2288	**M76559_at**	Neuronal DHP-sensitive, voltage-dependent, calcium channel alpha-2b subunit mRNA
2354	**M92287_at**	CCND3 Cyclin D3
2363	M93056_at	LEUKOCYTE ELASTASE INHIBITOR
2402	M96326_rna1_at	Azurocidin gene
2642	**U05259_rna1_at**	MB-1 gene
3252	**U46499_at**	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
4107	**X07743_at**	PLECKSTRIN
4196	**X17042_at**	PRG1 Proteoglycan 1, secretory granule
4328	**X59417_at**	PROTEASOME IOTA CHAIN
4366	**X61587_at**	ARHG Ras homolog gene family, member G (rho G)
4377	**U46499_at**	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
4847	**X95735_at**	Zyxin
5171	Z49194_at	OBF-1 mRNA for octamer binding factor 1
5501	**Z15115_at**	TOP2B Topoisomerase (DNA) II beta (180kD)
6041	**L09209_s_at**	APLP2 Amyloid beta (A4) precursor-like protein 2
6281	**M31211_s_at**	MYL1 Myosin light chain (alkali)
6855	**M31523_at**	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
1909	M29696_at	IL7R Interleukin 7 receptor
1953	M33195_at	Fc-epsilon-receptor gamma-chain mRNA
2335	M89957_at	IGB Immunoglobulin-associated beta (B29)

Table 4.5: Description des trente tops premiers gènes sélectionnés pour Leucémie, avec un nombre de fréquence complet (Freq = 30), à tavers 30 essais indépendants

4.5 Conclusions

En résumé, nous avons considéré dans ce chapitre l'application des méthodes de sélection de caractéristiques basées ensemble à l'identification de biomarqueurs. En effet, cette technique est une voie prometteuse pour une sélection plus stable et précise des gènes du cancer. Nous avons également proposé une approche massivement parallèle sur la base de méta-ensemble de filtres pour la découverte de biomarqueurs à partir des données de grande dimension. La méthode $MPME - FS$ est différente des autres méthodes de sélection de caractéristiques basées ensemble car elle effectue une sélection parallèle en deux étapes: la première au sein de chaque ensemble par l'agrégation des résultats des différents sélecteurs, et la deuxième étape représente l'agrégation des résultats de tous les ensembles en utilisant une deuxième fonction de consensus. Les biomarqueurs finaux sélectionnés sont utilisés pour construire un modèle de classification qui sera utilisé par la suite comme un outil efficace pour diagnostiquer les différentes sous-classes de cancer.

En outre, la méthode proposée est rapide et efficace en temps de calcul puisqu'elle est massivement parallèle et aucun algorithme d'apprentissage est utilisé dans le processus de sélection. Au lieu de cela, nous avons utilisé des filtres qui sont habituellement exploités quand le nombre de caractéristiques devient très grand en particulier pour les données à dimensions élevées. De toute évidence, $MPME-FS$ peut être effectuée en utilisant n'importe quel filtre de classement et appliquée à tout problème de sélection de caractéristique.

Les expériences sur cinq ensembles de données puce à ADN ont révélé que de bons résultats peuvent être obtenus par $MPME-FS$ en termes de performance de classification et de robustesse. L'analyse biologique des résultats montre que $MPME-FS$ offre une sélection de gènes hautement informatifs et qui ont des significations biologiques et sont aussi sélectionnés par d'autres approches.

Chapitre 5 Un ensemble de méta-heuristiques coopératives parallèles pour la sélection de gènes (ECPM-FS) dans la classification du cancer

Contenu du chapitre

5.1 Introduction

L'accumulation de preuves suggère que les méthodes d'ensemble et l'intelligence basée essaim sont deux solutions croissantes pour l'amélioration des algorithmes de sélection de caractéristiques. Dans ce travail, nous nous intéressons toujours aux méthodes d'ensemble qui ont été largement appliquées dans la bioinformatique pour améliorer la robustesse des signatures (Abeel *et al.*, 2010). La principale amélioration offerte par les méthodes d'ensemble est leur capacité à faire face à la malédiction de la dimensionnalité des données d'expression génique. Les méthodes d'ensemble proposées pour la sélection de caractéristiques afin de découvrir des biomarqueurs sont généralement à base de filtres. Au meilleur de nos connaissances, il n'y a pas de travail qui a proposé une méthode d'ensemble basée wrapper dans ce domaine.

L'étude actuelle présente une nouvelle méthode de sélection de caractéristiques d'ensemble basée wrapper, appelée ECPM-FS, qui permet d'identifier un nombre prédéfini de biomarqueurs à partir des données d'expression génique.

Comme les systèmes d'ensemble utilisent une procédure en deux étapes pour prendre des décisions, l'approche proposée (ECPM-FS) est conçue comme un ensemble de métaheuristiques parallèles et coopératives effectuée sur deux niveaux (Boucheham *et al.*, 2015a).
Pour ce faire, nous nous adaptons d'abord le modèle d'îles généralisée ("Generalized Islands Model" (GIM)) au problème de sélection de caractéristiques. Le GIM est un cadre général proposé récemment pour la coopération parallèle de plusieurs métaheuristiques basées population dans le domaine de l'optimisation (Izzo *et al.*, 2012). Ainsi, la génération des différentes parties de notre ensemble est basée sur une méthode hybride wrapper/filter que nous proposons, dans un premier temps, afin d'exploiter la force de plusieurs métaheuristiques au même temps pour faire la sélection, appelée (CPM-FS) (Boucheham *et al.*, 2015b). A ce stade, nous remplacons l'initialisation aléatoire dans les métaheuristiques traditionnelles par un mécanisme d'initialisation à base d'ensemble de filtres. En outre, afin de réparer les individus qui ont un nombre irrégulier de gènes sélectionnés, nous proposons un mécanisme de réparation à base de filtre qui peut également contribuer dans la sélection des gènes les plus informatives. Après un nombre prédéfini d'itérations et de migrations, chaque GIM sélectionne un sous-ensemble de meilleures caractéristiques.
La prochaine étape est l'agrégation de tous les sous-ensembles sélectionnés. Cette étape est réalisée par le biais d'un GIM en utilisant comme entrée un sous-ensemble réduit de caractéristiques (gènes). Ce dernier représente l'union de tous les gènes sélectionnés dans les différentes parties de l'ensemble, au cours de l'étape précédente. Ainsi, l'ECPM-FS permet de sélectionner un sous-ensemble plus robuste et précis de gènes, en particulier lors de l'utilisation de la méthode wrapper pour l'agrégation de différents sous-ensembles générés.

Le reste de ce chapitre est organisé comme suit. En premier, nous décrivons la méthode ECPM-FS proposée pour la découverte de biomarqueurs. Ensuite, les résultats expérimentaux sont discutés et on termine par des conclusions et des perspectives.

5.2 Principe de ECPM-FS

5.2.1 Le cadre général de ECPM-FS

Dans cette section, nous présentons une description détaillée de l'approche proposée pour la découverte de biomarqueurs à partir des données d'expression

génique. L'approche ECPM-FS proposée est conçue comme une méthode d'ensemble qui effectue la sélection de gènes en deux étapes.

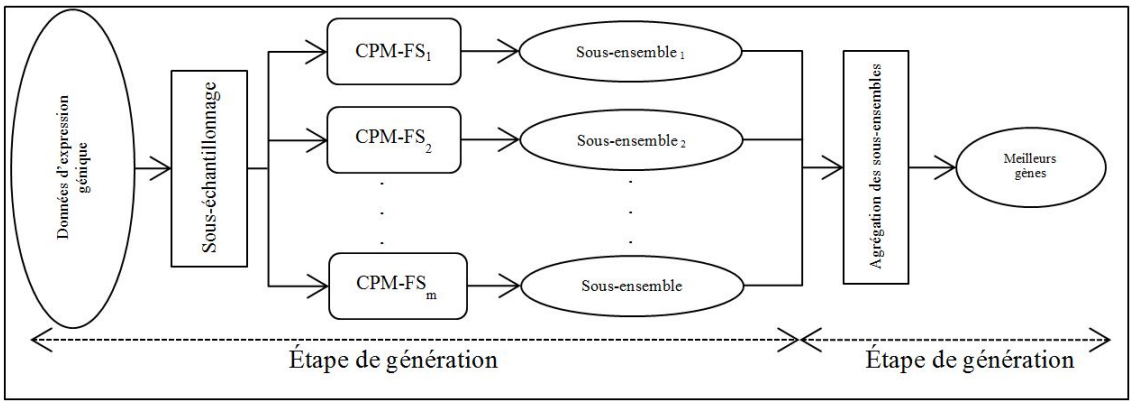


Figure 5.1: Organigramme décrivant le cadre général de l'ECPM-FS

Comme le montre la Figure 5.1, nous construisons d'abord les différentes parties de l'ensemble $sous-ensemble_1$, $sous-ensemble_2$ ... $sous-ensemble_m$ suivie d'une phase d'agrégation pour sélectionner le sous-ensemble final de meilleures gènes. Comme la construction des différentes parties de l'ensemble est l'étape la plus importante, nous utilisons un mécanisme hybride (wrapper/ filtre) pour effectuer la sélection (Boucheham *et al.*, 2015b). Ce dernier permet d'un coté de prendre en compte de la puissance de chaque gène séparément à l'aide d'un filtre dans le processus de sélection. En outre, il considère aussi la relation entre les gènes (force de groupe) en étudiant les différentes combinaisons de gènes à travers l'exploration de l'espace de solution par la méthode wrapper. Par ailleurs, la diversité au sein des parties de l'ensemble est garantie, car ils sont basées sur des méthodes stochastiques et aléatoires.

Afin d'assurer une meilleure exploration de l'espace des solutions, la sélection au sein de chaque partie de l'ensemble est basée sur l'Optimisation Coopérative et Parallèle (OCP). La coopération de plusieurs métaheuristiques inspirées de la nature représente une solution efficace pour éviter la convergence prématurée, tout en accélérant le processus de recherche (García-Nieto and Alba, 2012). Par conséquent, nous employons différentes métaheuristiques à base de populations telles que "Particules Swarm Optimization" (PSO), "Ant Colony Optimization" (ACO) et "Genetic algorithm" (GA) qui sont déployées en parallèle avec un mécanisme de migration des solutions (individus) (Martinez *et al.*, 2010).

Izzo et al. ont proposé le modèle d'îles généralisé (GIM), qui est un nouveau cadre pour mettre en œuvre l'OCP (Izzo *et al.*, 2012). Ce dernier emploie différents algorithmes d'optimisation sur différentes îles avec l'introduction d'un opérateur de migration. Selon Izzo et al, le GIM peut accélérer le temps de calcul de manière significative et augmente encore la précision de la prédiction par rapport au cas homogène dans plusieurs problèmes d'optimisation. Nous avons adapté le cadre général du GIM dans une méthode hybride wrapper/-filtre qui effectue la sélection de caractéristiques avec un nombre prédéfini à sélectionnés (Boucheham *et al.*, 2015b). Cette dernière constitue chaque partie de l'ensemble proposé avec une configuration spécifique.

Une fois la construction de toutes les parties de l'ensemble est atteinte, la première étape est finalisée par la sélection des meilleurs sous-ensembles de caractéristiques sur toutes les îles de chaque GIM. Ces sous-ensembles sont agrégés dans la deuxième étape par le biais d'une fonction de consensus à base wrapper, comme le montre la Figure 5.1.

5.2.2 L'étape de génération (CPM-FS)

Nous décrivons dans cette sous-section le modèle parallèle de métaheuristiques coopératives pour la sélection des caractéristiques qui constitue chaque partie de la méthode ECPM-FS. Il est basé sur le modèle d'île généralisée adapté à la sélection de caractéristiques (Izzo *et al.*, 2012). Afin d'atteindre l'objectif de la diversité fonctionnelle, nous utilisons à l'intérieur de chaque CPM-FS une configuration différente des métaheuristiques basées population (voir Figure 5.2) (Boucheham *et al.*, 2015b). Le processus de sélection est lancé par l'initialisation des populations étant les solutions initiales de chaque île. AL Gutiérrez et al, ont montré que les stratégies d'initialisation utilisées à l'intérieur des métaheuristiques se comportent différemment dans les différents types de problèmes, spécialement ceux qui ont un espace de recherche à grande dimension (Lanza *et al.*, 2011). Ainsi, les positions initiales des populations détermine et influence significativement la convergence des métaheuristiques.

En conséquence, nous proposons une nouvelle technique d'initialisation pour le problème de sélection de gènes qui se base sur un ensemble de filtres (Saeys *et al.*, 2008). Tous les sous-ensembles sélectionnés (S_1, S_2 ... S_N) à travers les différents filtres (F_1, F_2 ... F_N) au sein de l'ensemble représentent la population initiale dans chaque île comme le montre la Figure 5.2. Afin de construire les différentes parties de l'ensemble de filtres sur la base des sous-

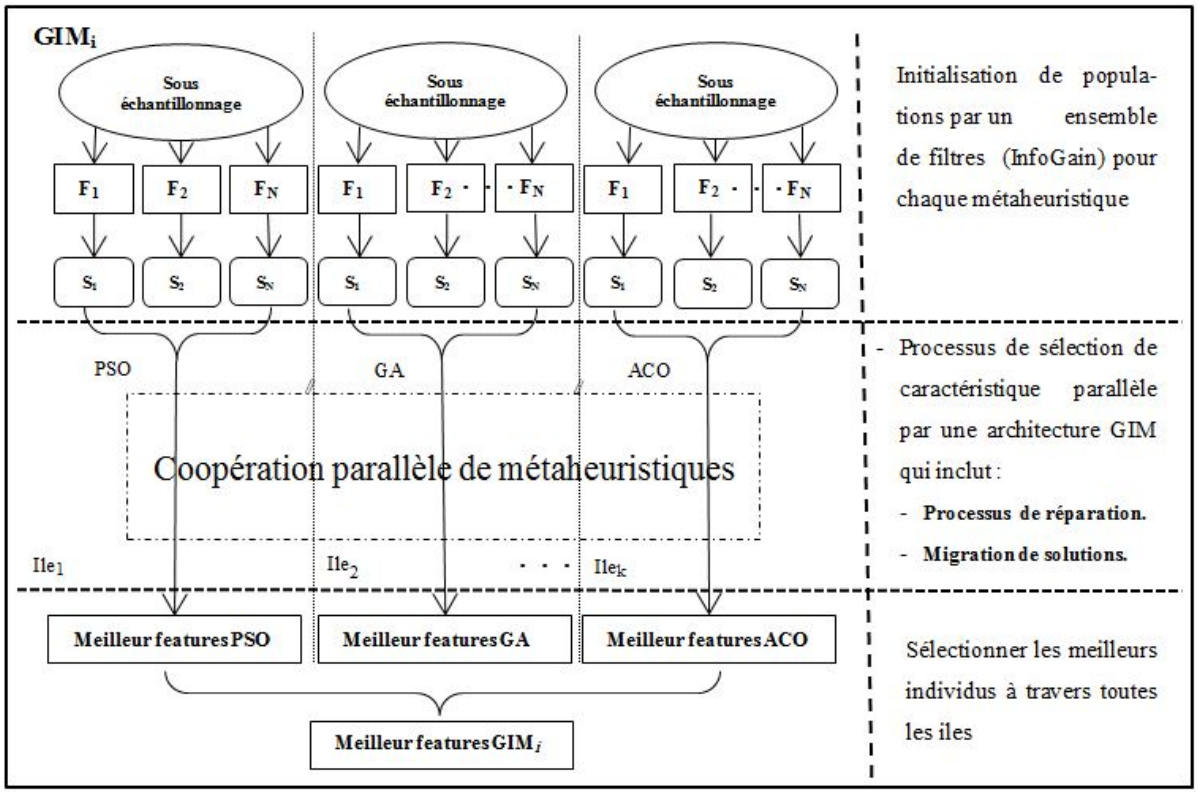


Figure 5.2: Modèle parallèle de chaque GIM-FS (CPM-FS)

échantillons déjà générés. Nous choisissons l'un des filtres les plus populaires et les plus réussis, qui est Information Gain (IG), car il est simple, rapide et convenable pour les méthodes d'ensemble (Quinlan, 2014). D'autre part, pour assurer la diversité de données entre les différentes populations, nous utilisons le partitionnement de données avec chevauchement qui permet de créer des ensembles de données réduits pour de nombreux sélecteurs au sein de l'ensemble de filtres. La perturbation des données consiste à générer des sous-échantillons en éliminant des échantillons de l'ensemble de données originales aléatoirement (Saeys *et al.*, 2008).

Par ailleurs, la migration des solutions se fait dans différents intervalles avec une communication asynchrone entre les îles en fonction d'une topologie entièrement connectée et une politique élitiste. Le choix approprié du mécanisme de migration peut empêcher les algorithmes d'optimisation de se retrouver dans un optimum local. En conséquence, nous pouvons obtenir de bons résultats ainsi que réduire le nombre d'itérations et la taille des populations à la fois. D'autre part, étant donné que le nombre de gènes sélectionnés

est prédéfini dans chaque solution, la fonction objective utilisée pour guider la recherche est seulement la précision donnée par le classificateur SVM en utilisant la procédure de la validation croisée (avec $k = 5$).
Une fois un critère de terminaison est satisfait dans toutes les îles, nous sélectionnons le sous-ensemble avec la meilleure valeur de fitness à travers tous les GIM-FS.

Processus de réparation: Afin d'appliquer les différentes métaheuristiques (PSO, GA, ACO ...) pour la sélection de gènes avec un nombre prédéfini à sélectionner n, certains ajustements doivent être effectués lors de la mise à jour des solutions de toutes les métaheuristiques. Constatant que les solutions sont représentées comme un vecteur binaire de longueur D (D est le nombre total de gènes). La valeur du i^{me} gène dans chaque vecteur de solution est soit 1 ou 0 pour indiquer si ce i^{me} gène est sélectionné ou non, respectivement.
L'application des opérateurs de mise à jour de chaque métaheuristique sur les solutions conduit au débordement du nombre de gènes sélectionnés dans chaque individu, il peut être inférieur ou supérieur du nombre désiré n (Boucheham *et al.*, 2015b).

Pour surmonter ce problème, nous proposons l'intégration d'un processus de réparation qui sera introduit pour chaque nouvelle solution qui ne satisfait pas à cette exigence (le nombre prédéfini de gènes). Le processus de réparation proposé est basé sur l'information donnée par le filtre Information Gain (IG). Par conséquent, on peut distinguer deux cas différents de réparation. Le premier est quand le nombre de gènes sélectionnés dans chaque individu est supérieur à n. Dans ce cas, certains gènes sélectionnés doivent être éliminés. A cet effet, nous utilisons le filtre IG pour classer tous les gènes sélectionnés et éliminer ceux qui sont moins classés. Le second cas est lorsque le nombre de gènes sélectionnés est inférieur à n. Dans ce cas, on ajoute au sous-ensemble les gènes les mieux classés à partir de tout l'ensemble de gènes initial avec l'exclusion de ceux déjà sélectionnés, comme le montre la Figure 5.3.

5.2.3 L'étape de consensus

Par analogie avec les méthodes d'ensemble, l'agrégation des sous-ensembles générés dans l'étape précédente est un élément clé dans notre méthode ECPM-FS. L'objectif principal consiste à agréger M sous-ensembles de caractéristiques pour obtenir un sous-ensemble final contenant les caractéristiques les

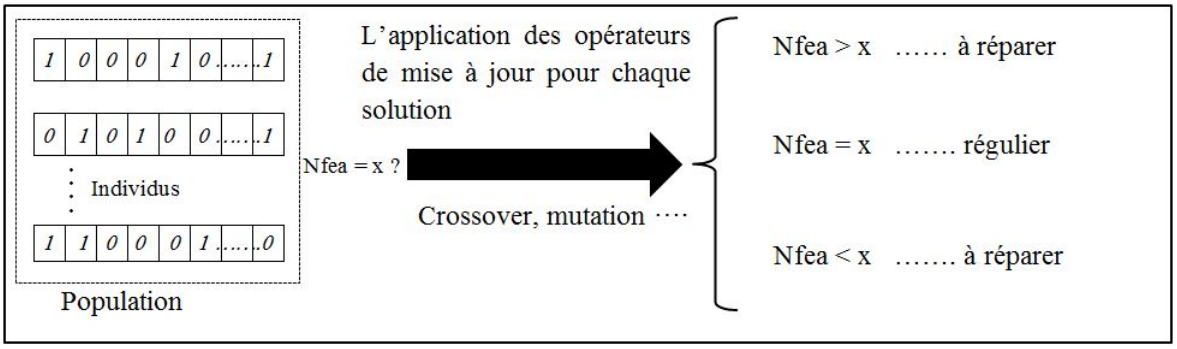


Figure 5.3: Le débordement du nombre sélectionné de biomarqueurs dans le processus de sélection du CPM-FS

plus représentatives (gènes). L'utilisation des fonctions de consensus qui sont basées sur l'étude de chaque gène séparément comme compter les caractéristiques les plus fréquentes n'est pas la solution la plus efficace dans notre cas.
Ainsi, une meilleure fonction de consensus doit prendre en compte les spécificités du problème biologique à traiter. Dans le problème de la découverte de biomarqueurs, il est préférable que la fonction de consensus préserve la relation entre les gènes (force du groupe) qui travaillent ensemble pour atteindre un objectif commun dans la cellule.

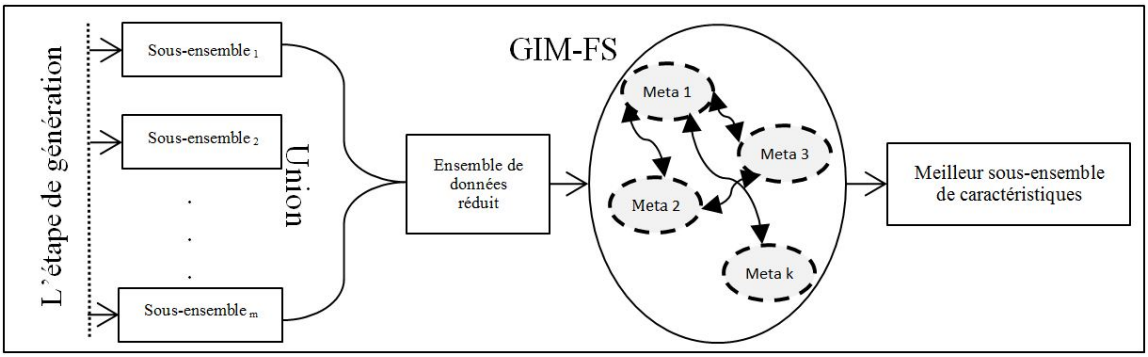


Figure 5.4: Processus de consensus: deuxième étape dans ECPM-FS

Afin d'atteindre cet objectif, nous proposons une fonction de consensus basée wrapper pour agréger les différents sous-ensembles sélectionnés dans l'étape précédente. L'agrégation se fait en effectuant un GIM-FS à partir d'un ensemble réduit de gènes. Ce dernier représente l'union de tous les gènes appartenant aux sous-ensembles sélectionnés, comme le montre la Figure 5.4.

La même méthode CPM-FS présentée ci-dessus est utilisée pour établir un consensus entre les parties de l'ensemble, et sélectionne le sous-ensemble de meilleures biomarqueurs.
Enfin, les biomarqueurs sélectionnés sont utilisés pour construire un modèle de classification qui aide à prendre les décisions appropriées concernant le traitement des maladies en question. Cela peut fournir aux patients un meilleur traitement, en particulier lorsque la maladie sera détectée dans ses débuts.

5.3 Résulats et discussions

5.3.1 Paramètres et jeux de données (GED)

Dans cette section, nous présentons l'étude empirique menée afin d'évaluer la performance de l'approche proposée et de la comparer à d'autres méthodes dans la littérature. Notre étude se concentre sur la sélection de caractéristiques supervisée, puisque plusieurs ensembles de données de puces à ADN ont des valeurs de classe qui sont utiles pour la prédiction. Nous avons mis en place notre méthode en utilisant la bibliothèque "Parallel Computing Toolbox" (PCT) de MATLAB®. Afin de tester l'efficacité de l'ECPM-FS proposé, douze ensembles de données de puces à ADN sont utilisés. Le Tableau 5.1 résume les détails de ces ensembles de données.

Données	#Gènes	#Classes	#Echantillons
9_tumors	5.726	9	60
11_tumors	12.533	11	174
Prostate_Tumor	10.509	2	102
Colon	2.000	2	62
Leukemia	7.129	2	72
Ovarian	15.154	2	253
Leukemia1	5.327	3	72
Leukemia2	11.225	3	72
DLBCL	5.469	2	77
SRBCT	2.308	4	83
Brain_Tumor1	5.920	5	90
Brain_Tumor2	10.367	4	50

Table 5.1: Les caractéristiques des différents ensembles de données utilisés

L'ECPM-FS proposé se compose d'un ensemble de quatre GIM-FS (CPM-

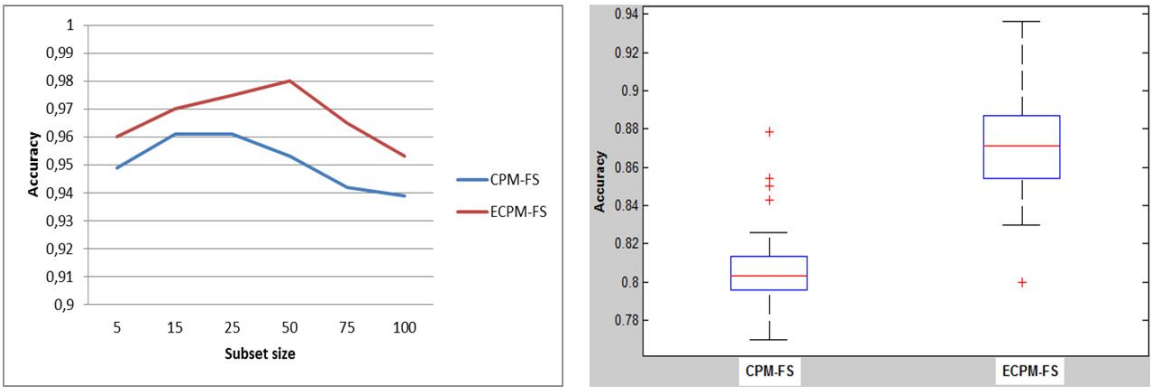
FS) dont chacun est composé de trois îles. Chaque île effectue la sélection par une métaheuristique basée populations modifiée comme présenté dans la section précédente. Dans ce but, nous utilisons trois métaheuristiques bien connues à savoir PSO, GA (Alba *et al.*, 2007) et ACO (Huang, 2009). Les paramètres expérimentaux des différents GIM-FS sont donnés dans le Tableau 2.

Politique de sélection	stratégie élitiste
Politique de recombinaison	Remplacement des mauvaises solutions
La topologie	bidirectionnelle entièrement connectée
Communication	asynchrone initié par la source
Nombre maximum d'itérations	250
Intervalle de migration	50
Fonction de fitness	précision du classificateur SVM avec 5-fold cross-validation
Taille de population/essaim	30
c1 , c2 du PSO	2
Probabilité de croisement	0.8
Probabilité de mutation	0.05

Table 5.2: Réglage de paramètres du $ECPM - FS$

5.3.2 Résultats

Pour mesurer la contribution de l'utilisation d'un ensemble de GIM-FS avec la coopération des métaheuristiques, nous comparons tout d'abord la capacité prédictive en terme de précision de classification de l'approche proposée basée ensemble ECPM-FS (Boucheham *et al.*, 2015a) avec la méthode simple CPM-FS (Boucheham *et al.*, 2015b).
La Figure 5.5.a, rapporte les valeurs moyennes de précision des deux méthodes en fonction de la taille du sous-ensemble à sélectionner, effectuées sur le cancer du côlon. Dans cette figure, on peut clairement observer que l'ECPM-FS peut atteindre une précision de classification plus élevée qu'un simple CPM-FS, indépendamment de la taille de sous-ensemble sélectionné. En outre, la Figure 5.5.b montre les boxplots obtenus à travers 40 exécutions sur le jeu de donné "9 tumeurs".
Etant donné que les autres études mentionnées dans le Tableau 5.3 ne sont pas suffisamment dotées d'informations et de configurations de leurs méthodes, une comparaison complète ne peut pas se faire. Ainsi, nous effectuons une



(a) (b)

Figure 5.5: Comparaison de performances: (a) comparaison des deux méthodes ECPM-FS et CPM-FS en fonction de la taille des sous-ensembles séléctionnés à travers l'ensemble de données "Colon", (b) les Boxplots de ECPM-FS et CPM-FS à travers l'ensemble de données "9_tumeurs"

comparaison basée sur les résultats mentionnés dans certaines études récentes dans la littérature. Le Tableau 5.3 résume les résultats obtenus par nos méthodes proposées dans le cadre de cette thèse et six d'autres approches en termes de précision et de nombre de biomarqueurs sélectionnés (IBPSO(Martinez *et al.*, 2010), cuPSO (Martinez *et al.*, 2010), CBBBOFS(Yazdani *et al.*, 2013), CIBBOFS(Yazdani *et al.*, 2013), PMSO(García-Nieto and Alba, 2012) et EFS(Saeys *et al.*, 2008)). Afin de tirer des conclusions statistiquement significatives, 30 essais indépendants ont été réalisés pour chaque ensemble de données. Les valeurs moyennes obtenues sont consignées dans le Tableau 5.3. On peut voir à partir de ce tableau que l'ECPM-FS donne une amélioration significative par rapport à la version simple CPM-FS et surpasse les autres approches dans presque tous les jeux de données.

Par ailleurs, afin d'analyser la robustesse des signatures sélectionnées à travers ECPM-FS, nous avons évalué la similitude entre les sous-ensembles sélectionnés dans différents essais indépendants par notre méthode. La stabilité globale S_{tot} comme définie précédemment, et calculée en utilisant deux mesures de similarité:

$$Indice \quad de \quad Jaccard = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} \tag{5.1}$$

$$Indice \quad de \quad Kuncheva = \frac{|f_i \cap f_j| - \frac{S}{N}}{S - \frac{S}{N}} \tag{5.2}$$

	Nos approches		Autres approches dans la littérature					
	CPM-FS	**ECPM-FS**	IBPSO	cuPSO	CBBBOFS	CIBBOFS	PMSO	EFS
9_tumors	0.804 (25)	**0.8757 (25)**	0.783(1280)	0.85(149)	76.5(25)	0.7156 (25)	/	/
11_Tumors	0.85 (25)	0.874 (25)	0.931(2948)	0.936(535)	**0.8879 (25)**	0.8549 (25)	/	/
Prtate_Tumor	0.986 (25)	0.992 (25)	0.921(1294)	0.99(4)	**0.993 (25)**	0.9718 (25)	/	/
Colon	0.961 (25)	**0.975 (25)**	/	/	/	/	0.942(20)	0.87(20)
Leukemia	1(25)	**1(25)**	/	/	/	/	0.981(20)	0.98(71)
Ovarian	1(25)	**1(25)**	/	/	/	/		0.97(151)
Leukemia1	1(25)	**1(25)**	1(1034)	**1(5)**	/	/	/	/
Leukemia2	1(25)	**1(25)**	1(1292)	**1(5)**	**1(25)**	0.995 (25)	/	/
Brain_Tumor1	0.964 (25)	0.976 (25)	0.944 (754)	**0.977 (12)**	/	/	/	/
Brain_Tumor2	0.974(25)	**1(25)**	0.94(1197)	0.88 (161)	/	/	/	/
DLBCL	1(25)	**1(25)**	1(1042)	**1(3)**	/	/	/	/
SRBCT	1(25)	**1(25)**	1(431)	**1(5)**	/	/	/	/

Table 5.3: Résultat de la comparaison basée sur la précision de classification et le nombre de biomarqueurs sélectionnés (#)

La Figure 5.6 résume les résultats moyens de la robustesse de ECPM-FS à travers les différents jeux de données, au moyen de deux mesures de similarité: l'indice de Jaccard et de Kuncheva.

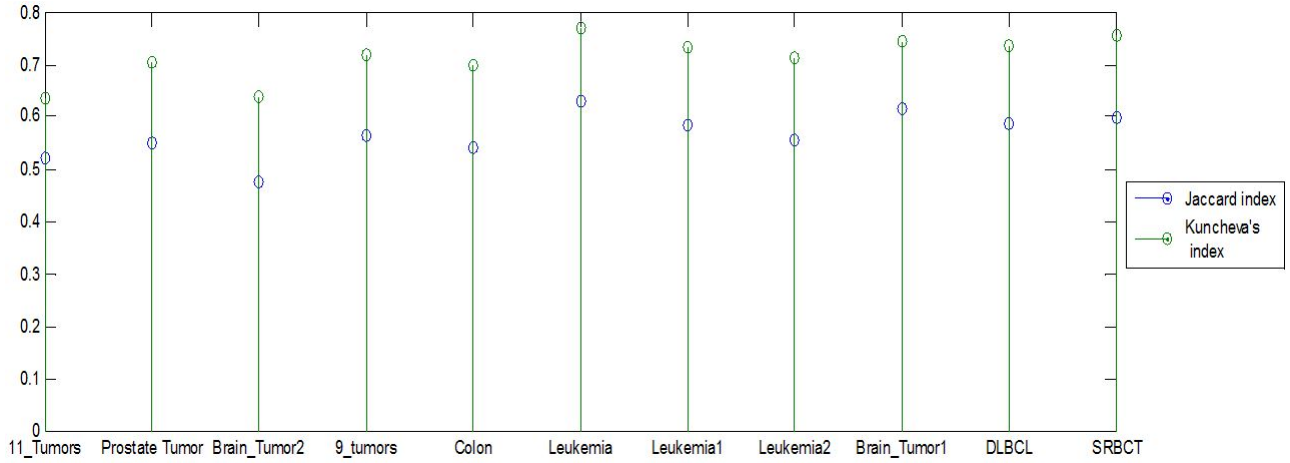


Figure 5.6: Similarités moyennes en fonction de deux indices: Jaccard et Kuncheva

5.3.3 Discussions

D'après les résultats donnés ci-dessus, On peut clairement conclure que l'ECPM-FS donne de bons résultats et surpasse toutes les méthodes dans

la littérature, en terme de précision de classification. Même si les méthodes wrapper ne sont pas robustes en raison de leur comportement stochastique. En particulier dans le cas de données de puces à ADN, qui comprennent des gènes redondants dans des espaces de grande dimension. Les résultats de robustesse moyens présentés dans la Figure 5.6 indiquent que la méthode ECPM-FS est relativement robuste sur tous les ensembles de données. Les mécanismes à base de filtres intégrés dans l'initialisation des populations comme dans le processus de réparation permettent de minimiser le degré d'incertitude dans la procédure de recherche dans l'espace de solution et d'éviter la sélection des gènes redondants. De l'autre côté, la partie ensemble permet d'établir un consensus entre plusieurs sélections menant à une sélection plus robuste et précise. Par conséquent, l'ECPM-FS permet principalement d'établir un compromis entre la robustesse et les performances élevées de prédiction.

Prostate Tumor	Indx	4823	8220	7652	6105	7451	7515	9949	10130	5815	9138	10125	7529	8765	120	181
	Freq	40	40	39	35	35	35	35	35	34	32	31	29	29	28	28
11_Tumors	Indx	542	581	7093	8037	9706	2768	4787	6139	7651	7735	7789	8168	11764	1808	2962
	Freq	40	40	40	40	40	38	38	38	38	38	38	38	38	36	36
9_tumors	Indx	80	1755	5032	5183	1361	1777	3372	4138	5147	15	1916	4604	3283	4996	1354
	Freq	40	40	40	40	39	39	38	38	38	37	37	37	36	36	35
Brain Tumor1	Indx	3586	5361	1183	4688	2610	2642	1497	1595	1965	5453	2532	5391	2478	244	505
	Freq	40	40	39	39	39	39	37	34	34	34	33	32	31	30	30
Brain Tumor2	Indx	276	423	687	915	1245	210	703	2313	5847	1696	10337	2846	4863	9568	9801
	Freq	40	40	40	40	40	39	39	39	39	38	38	37	37	37	37
Colon	Indx	1635	138	1060	1110	493	377	897	1263	1365	241	249	267	1549	1960	365
	Freq	40	39	39	39	38	37	37	37	37	36	36	36	35	35	34
DLBCL	Indx	1259	3257	3942	2164	226	409	1600	717	1670	773	856	3127	874	1122	5250
	Freq	40	40	40	39	34	33	31	30	28	27	27	27	26	26	26
Leukemia	Indx	4951	2354	4366	1834	4328	6855	1144	5501	1882	1902	2121	2642	6041	758	1685
	Freq	40	36	35	33	32	32	30	30	29	29	29	28	28	27	27
Leukemia1	Indx	1999	5142	618	1770	4009	1271	2350	3549	1611	3076	728	1426	4307	4688	3518
	Freq	40	39	38	37	37	35	34	34	33	33	32	32	32	32	31
Leukemia2	Indx	225	4992	7050	6746	7545	10355	832	4915	10174	1108	2079	5657	6118	4845	6720
	Freq	40	39	39	35	35	35	34	34	34	33	33	33	33	32	32
SRBCT	Indx	174	1389	1	1073	477	1613	368	1932	338	819	1315	107	108	1263	1700
	Freq	40	39	38	38	37	37	37	37	36	36	35	33	33	33	33
Ovarian	Indx	2313	1599	1679	1684	1688	2237	2240	181	1682	2238	2191	2193	1675	1689	1735
	Freq	40	39	39	39	39	39	39	38	38	38	37	37	36	36	36

Table 5.4: Top quinze gène séléctionnés (indice de gène (Index) et leur fréquence (Freq)) à travers les douze ensembles de données sur 40 exécutions

Enfin, nous fournissons les signatures sélectionnées en utilisant l'ECPM-FS à travers 40 essais indépendants, comme on le voit sur le Tableau 5.4. Les gènes appartenant aux quinze premiers gènes sélectionnés ont été choisis avec une fréquence très élevée, ce qui confirme la capacité de notre méthode

à sélectionner une signature robuste. Nous fournissons également aux experts biologistes et aux cliniciens l'interprétation biologique de la signature sélectionnée à partir de l'ensemble de données SRBCT, comme l'indique le Tableau 5.5.

Indice de Gène	Numéro d'accession	"Hugo name"	Description de Gène
174	4771	NF2	neurofibromin 2 (bilateral acoustic neuroma)
1389	2217	FCGRT	Fc fragment of IgG, receptor, transporter, alpha
1	1495	CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa
1073	5045	FURIN	furin (paired basic amino acid cleaving enzyme)
477	1942	EFNA1	ephrin-A1
1613	8991	SELENBP1	selenium binding protein 1
368	7088	TLE1	transducin-like enhancer of split 2, homolog of Drosophila E(sp1)
1902	3159	HMGA1	high mobility group AT-hook 1
1932	2737	GLI3	GLI family zinc finger 3
338	4330	MN1	meningioma (disrupted in balanced translocation) 1
819	6258	RXRG	retinoid X receptor, gamma
1315	2619	GAS1	growth arrest-specific 1
107	7295	TXN	thioredoxin
1263	3316	HSPB2	heat shock 27kD protein 2
1700	2275	FHL3	ESTs, Moderately similar to skeletal muscle LIM-protein FHL3 [H.sapiens]

Table 5.5: Top quinze gènes sélectionnés du jeu de données "SRBCT"

5.4 Conclusions

Nous avons abordé dans ce chapitre le problème de la sélection de biomarqueurs significatifs à partir des données d'expression génique de grande dimension en génomique. Nous proposons une procédure en deux étapes basées sur un ensemble de métaheuristiques coopératives et parallèles (ECPM-FS) pour la découverte d'un nombre prédéterminé de biomarqueurs utiles. Les meilleurs sous-ensembles de gènes sont agrégés à travers les îles et sont de nouveau traitées par un GIM-FS. Les gènes sélectionnés peuvent être ensuite utilisés pour une analyse ultérieure, comme la classification des types de cancer. Les résultats expérimentaux sur douze jeux de données de puces à ADN ont montré que notre approche fonctionne mieux que d'autres méthodes existantes dans la littérature récente. En outre, l'approche proposée peut être facilement étendue à tout problème de sélection de caractéristiques et adaptée pour les grands ensembles de données.

Chapitre 6

Approche integrative pour la prediction des ARN non-codants. application sur les ARNs Piwi

Contenu du chapitre

6.1 Introduction

Les ARN non-codants (ARNnc) sont des biomarqueurs moléculaires qui jouent des rôles importants dans diverses activités cellulaires et sont étroitement associés au cancer et d'autres maladies complexes. Ce qui a fait leur identification un problème critique dans les recherches biologiques (Esteller, 2011). Avec le développement des nouvelles générations de technologies de séquençage (NGS), les biologistes peuvent maintenant accéder à d'énormes volumes de données de séquençage (par exemple d'ARN-Seq). Exploiter cette quantité de données nécessite des outils de calcul afin d'arriver à l'identification de potentiels ARNnc, suivie d'une validation expérimentale des candidats.

Plusieurs types d'ARNnc ont été découverts, beaucoup d'entre eux récemment, ce qui suggère que d'autres seront certainement découverts au cours des prochaines années. La majorité de ces ARNnc se caractérisent par une structure secondaire spécifique et / ou des motifs spécifiques. Ainsi, presque tous les outils existants pour la prédiction d'ARNnc sont basés sur ces caractéristiques classiques, et sont conçus spécifiquement pour un seul type d'ARNnc

(Soldà *et al.*, 2009). Par exemple, un grand nombre d'outils ont été développés pour les microARN, une classe largement étudiée d'ARNnc. Parmi ces outils, nous pouvons citer miRNAFold (Tempel and Tahi, 2012) et miRBoost (Tran *et al.*, 2015). Les caractéristiques de structure secondaire et de séquence ne sont pas suffisamment discriminantes pour prédire certaines classes d'ARNnc.

Dans le cadre de cette thèse, nous présentons une approche globale et intégrative pour la prédiction des ARNnc, en tenant compte de nombreuses caractéristiques récemment découvertes, qui pourraient être utilisées pour les caractériser. À cette fin, nous réalisons une étude globale sur ce qui peut caractériser un ARNnc de point de vue génomique et épigénomique. En effet, un ARNnc peut être caractérisé par (i) sa séquence, (ii) sa structure secondaire possible, mais également (iii) ses positions sur la chromatine, (iv) ses positions relatives à des séquences et / ou motifs structuraux qui peuvent se produire au niveau des extrémités 5' et / ou 3',(v) l'apparition éventuelle dans des clusters, et (vi) la possibilité d'interaction avec des séquences cibles spécifiques.

Nous développons ensuite un outil générique, appelé « IncRId », basé sur la méthode d'apprentissage à noyaux multiples (MKL) (Jain *et al.*, 2012). Cette méthode, qui combine plusieurs noyaux représentant différents types de caractéristiques, permet de prendre en compte leur hétérogénéité. En effet, nous définissons un ensemble de noyaux génériques qui pourraient être directement utilisés ou adaptés en fonction de la classe considérée d'ARNnc. Grâce au cadre orienté objet que nous proposons, notre outil est modulaire et facilement extensible et modifiable, et permet de tester chaque noyau séparément afin de percevoir la conservation de certaines caractéristiques biologiques à travers les espèces.

Dans une deuxième partie, nous présentons un exemple d'application de notre outil générique IncRId: la prédiction des ARN piwi (ARNpi). Les ARNpi sont la plus grande classe hétérogène de la famille des petits ARN non-codants. Ils manquent de motifs clairs de structure secondaire et aussi la conservation de la séquence primaire, ce qui rend leur prédiction une tâche très difficile. Nous avons fait une étude approfondie sur cette molécule, ce qui nous a permis de déduire plusieurs caractéristiques de différents types, qui sont mises en œuvre dans douze noyaux différents. L'outil résultant, appelé IpiRId, montre la surperformance et les avantages des approches intégratives dans la prédiction des ARNpi, par rapport à tous les autres outils

existants, à savoir piRNApredictor (Zhang *et al.*, 2011a), Piano (Wang *et al.*, 2014a), Pibomd (Liu *et al.*, 2014). et piRPred (Brayet *et al.*, 2014). IpiRId donne plus de 90% de précision pour chacune des trois espèces étudiées: l'humain, la souris et la mouche. Plus important, les résultats de prédiction sont homogènes pour toutes les espèces, ce qui n'est pas le cas pour les autres méthodologies et outils. Le logiciel IpiRId et le serveur web de notre outils sont disponibles gratuitement pour les utilisateurs universitaires sur: `https://tanuki.ibisc.univ-evry.fr/evryrna/IpiRId` .

6.2 Principe

Nous proposons ici une méthodologie générale et intégrative pour la prédiction des ARNnc basée sur l'apprentissage automatique supervisé qui peut (i) être appliqué pour prédire tout type d'ARNnc et (ii) intégrer différentes types de caractéristiques qui caractérisent un ou plusieurs types d'ARNnc.

6.2.1 Les caractéristiques des ARNnc

Dans ce qui suit, nous examinons et proposons une taxonomie des principaux aspects caractérisants les ARNnc. Ainsi, nous avons extrait de nombreuses caractéristiques de la littérature qui pourraient être prises en compte dans des approches computationnelles et contribuent dans leur prédiction.

Séquence

Plusieurs caractéristiques peuvent être calculées sur la séquence comme l'occurrence, la fréquence ou la position d'un motif dans la séquence de l'ARN. La composition en nucléotide peut aussi être utilisée pour caractériser une classe d'ARNs non-codants. En effet, les micros-ARNs matures peuvent être prédits en se basant sur leur premier nucléotide, la composition des 8 premiers nucléotides[1] et la composition en A/U ou G/C (Menor *et al.*, 2015). Concernant les ARNt, leur prédiction peut être basée sur la présence de motifs consensus ou en cherchant des nucléotides spécifiques à certaines positions de l'ARN (Laslett and Canback, 2004).
De plus, les snoARN ont un pourcentage de GC particulier et peuvent contenir des "boîtes" C, D H ou ACA qui peuvent être identifiées par des motifs consensus (Hertel *et al.*, 2008). Enfin les ARNpi sont souvent caractérisés par des nucléotides à certaines positions comme une uridine en première position ou une adénine en position 10 de l'ARN. Bien qu'ils ne présentent pas de

[1] nucléotides intervenant dans l'interaction avec un ARNm cible

motifs consensus identifiés au sein de leur séquence (Le Thomas *et al.*, 2014; Menor *et al.*, 2015), les ARNpi et miARNs matures peuvent être prédits en calculant la fréquence de tous les k-mers possibles puis en sélectionnant les plus discriminants (Zhang *et al.*, 2011a; Menor *et al.*, 2015).

Structure

La structure d'un ARN peut jouer un rôle primordial, dans sa prise en charge/interaction avec d'autres protéines et classes d'ARNs. Certains d'entre eux ont des structures particulières qui peuvent être discriminantes, en prédisant leur repliement à partir de la séquence. Par exemple, les structures secondaires d'ARNt prenant généralement la forme d'une feuille de trèfle (Laslett and Canback, 2004). Les SnoARN ont également une structure secondaire particulière, qui est utilisée pour leur prédiction (Hertel *et al.*, 2008). Autres ARNnc, comme les ARNpi, ne disposent pas d'une structure secondaire. Certains d'autres, comme les microARN (ARNmi), sont contenus dans des précurseurs (pré-ARNmi) qui se replient dans une structure en épingle à cheveux. Par conséquent, de nombreux outils développés pour prédire les pré-ARNmi sont essentiellement basés sur cette caractéristique.

Cluster

Des études récentes suggèrent que de nombreux ARNnc se regroupent en clusters sur le génome. Par exemple, les ARNpi présentent cette caractéristique et quelques outils l'utilisent afin de les prédire (Jung *et al.*, 2014; Brayet *et al.*, 2014). Aussi, les microARN apparaissent en clusters polycistroniques de plus d'un ARNmi et cette caractéristique a été utilisée d'une manière computationnelle pour leur identification (Washietl *et al.*, 2012; Chan *et al.*, 2012). En outre, d'autres observations suggèrent que les snoARN peuvent également former des clusters sur le génome (Makarova and Kramerov, 2007).

Interaction/Liaison avec des ARNs cibles

De nombreuses observations ont révélé que certains ARNnc interagissent avec des séquences cibles afin de jouer un rôle spécifique dans la cellule. Par exemple, les ARNpi sont connus pour inhiber l'action des éléments transposables (TE) (Weick and Miska, 2014). Cette caractéristique a été utilisée par (Wang *et al.*, 2014a) pour prédire les ARNpi en évaluant leur alignement / appariement avec les TEs. Aussi, les microARN matures peuvent se lient avec des ARNm (Menor *et al.*, 2015).

Éléments spécifiques autour de la séquence

Un autre aspect important qui peut être étudié est la présence ou l'absence de certains éléments autour de la séquence d'ARN sur le génome. Par exemple, les régions proximales des microARN contiennent rarement d'autres petits ARN non-Mir (Washietl *et al.*, 2012) et certaines études proposent de prédire les microARN en recherchant en amont des motifs consensus (Ohler *et al.*, 2004). Chez la souris l'expression de certains ARNpi nécessite le facteur de transcription A-Myb qui se fixe à l'ADN (Li *et al.*, 2013). La présence de ce motif de liaison en amont de la séquence pourrait être une caractéristique utilisée pour la prédiction. De plus certains clusters d'ARNpi sont entourés par des éléments transposables (Betel *et al.*, 2007).

Reliées à l'état de la chromatine

Une catégorie importante de caractéristiques qui peuvent significativement améliorer la prédiction d'ARNnc concerne les informations relatives à leur expression. Dans le noyau des eucaryotes le génome est organisé en trois dimensions dans l'espace sous la forme d'un complexe nucléoprotéique, appelé la chromatine. Celle-ci permet de compacter l'ADN et joue notamment un rôle dans la régulation de la transcription. Ce complexe est composé de protéines appelées histones sur lesquelles s'enroule l'ADN. Ces histones peuvent être méthylés et cela va influer sur le compactage de la chromatine ou sur le recrutement de facteurs de transcription.
En effet, chez la mouche, les clusters d'ARNpi sont souvent entourés de l'histone 3 tri-méthylé sur sa lysine 9 (H3K9me3) qui va recruter la protéine Rhino et ainsi permettre la transcription du cluster (Yu *et al.*, 2015). Considérer des caractéristiques telles que la distance à la plus proche méthylation d'une histone particulière peut être intéressant pour discriminer une classe d'ARN non codant. D'autres études sur les mêmes espèces ont montré que les clusters d'ARNpi sont souvent trouvés dans les régions de l'hétérochromatine centromériques et télomériques (Yamanaka *et al.*, 2014).

6.2.2 Le cadre MKL

L'utilisation d'un classificateur avec de nombreuses caractéristiques hétérogènes exige différents types de données provenant de différentes sources, ce qui ne permet pas l'utilisation d'une méthode de classification supervisée simple. Pour surmonter ce problème, nous proposons l'utilisation de l'apprentissage à noyaux multiple (MKL), une approche qui permet de combiner différents noyaux en ajustant automatiquement leurs poids (Gönen and Alpaydın,

2011). Nous construisons donc plusieurs noyaux indépendants représentant différents types d'informations (biologiquement ou de calcul), et puis les combiner afin d'effectuer une classification binaire en utilisant le classificateur SVM ("Support Vector Machine"). Nous utilisons le logiciel SPG-GMKL qui emploie le "spectral projected gradient descent" comme optimiseur afin de trouver la combinaison optimale de noyaux (Jain *et al.*, 2012). Dans ce travail, nous choisissons un noyau gaussien qui est un noyau universel pour la représentation de caractéristiques. Il se compose d'une matrice carrée de similarité de taille $(N * N)$, N étant la taille de l'ensemble de données d'apprentissage (échantillons: positifs et négatifs). Soient x et y deux vecteurs de caractéristiques ou des matrices représentant deux séquences. Le produit scalaire de x et y dans l'espace de caractéristique est donné par l'équation suivante:

$$k(x, y) = exp^{-\gamma||x-y||^2} \tag{6.1}$$

La distance euclidienne est utilisée si x,y sont des vecteurs ; si x,y sont des matrices (cas de la caractéristique de cluster) la distance de Frobenius est utilisée.

Afin de calculer d'une manière automatique le paramètre de noyau γ le plus approprié, une méthode souvent utilisée est la recherche par grille qui consiste à rechercher plusieurs valeurs de gamma et à tester chacune d'entre elles avec un classificateur. Le gamma permettant la meilleure classification parmi ceux testés sera alors sélectionné. Le problème de cette démarche est qu'elle est couteuse en temps, car pour chaque évaluation il faut lancer un classificateur.

Nous proposons ici une autre méthode d'évaluation plus rapide en se basant sur la distance inter-cluster (Wu and Wang, 2009). En effet, nous avons deux classes ou clusters : positifs et négatifs. Si la distance entre ceux-ci est grande, alors trouver une séparation entre les deux (SVMs) sera facilité. D'abord l'heuristique de Jaakkola est utilisée pour calculer la valeur initiale du γ de la façon suivante (Jaakkola *et al.*, 1999):

$$\gamma JAAK = 1/(2median(distMat)^2) \tag{6.2}$$

Ensuite, nous cherchons des solutions possibles de γ:

$$\gamma = exp(i) * \gamma JAAK \tag{6.3}$$

avec i un nombre entier dans $[-4; 4]$. Chacune d'entre elles est évaluée en calculant la distance inter-cluster entre les séquences positives et négatives.

Enfin, nous choisissons celle qui donne la distance la plus élevée, ce qui conduira à une meilleure classification (Wu and Wang, 2009). La distance inter-cluster est calculée comme suit (Wu and Wang, 2009):

$$\delta(X_+, X_-) = \frac{1}{l_+ + l_-} \left(\sum_{x_+ \in X_+} d(x_+, x_-^-) + \sum_{x_- \in X_-} d(x_-, x_+^-) \right) \tag{6.4}$$

où X_+ et X_- sont les séquences positives et négatives, l_+ et l_- sont leurs tailles correspondantes, et X_+^- et X_-^- sont les moyennes des classes X_+ et X_- .

6.2.3 Les principales classes de noyaux et le cadre orienté objet

Nous avons développé un cadre orienté objet, appelé IncRId, implémenté en Java, qui se compose de différentes classes et sous-classes représentant différents noyaux (kernels). La Figure 6.1 donne l'architecture générale de notre cadre et les différentes classes que nous avons définies. Certaines classes sont abstraites (bleu), car ils ne correspondent pas aux noyaux mis en œuvre, mais leur définition nous permet de construire une meilleure structuration hiérarchique des différents noyaux. En outre, certaines classes de kernels peuvent être instanciées pour tout type d'ARNnc (marron), et d'autres doivent être spécialisées, selon le type considéré d'ARNnc (vert). Dans ce qui suit, nous donnons une brève description des principales classes de kernels qui pourrait être instanciées directement pour tout ARNnc.

Motifs spécifiques dans de la séquence « Specific motifs inside »

Cette classe de kernels représente les caractéristiques correspondantes à la présence / absence d'un ensemble de motifs à des positions spécifiques dans la séquence. En conséquence, on construit un vecteur binaire de dimension N contenant les informations sur la présence ou l'absence de chaque motif dans la séquence, où N est le nombre de motifs.

Les Motifs "K-mer" « K-mer motifs »

Les "K-mers" sont largement appliqués pour l'analyse de séquence et se réfèrent à des k-tuplets spécifiques de nucléotides. Nous déterminons les k-mers les plus discriminants en générant un grand nombre d'entre eux et en effectuant ensuite une sélection sur cet ensemble élargi de k-mers pour identifier ceux qui sont les plus discriminants. Afin d'atteindre des performances

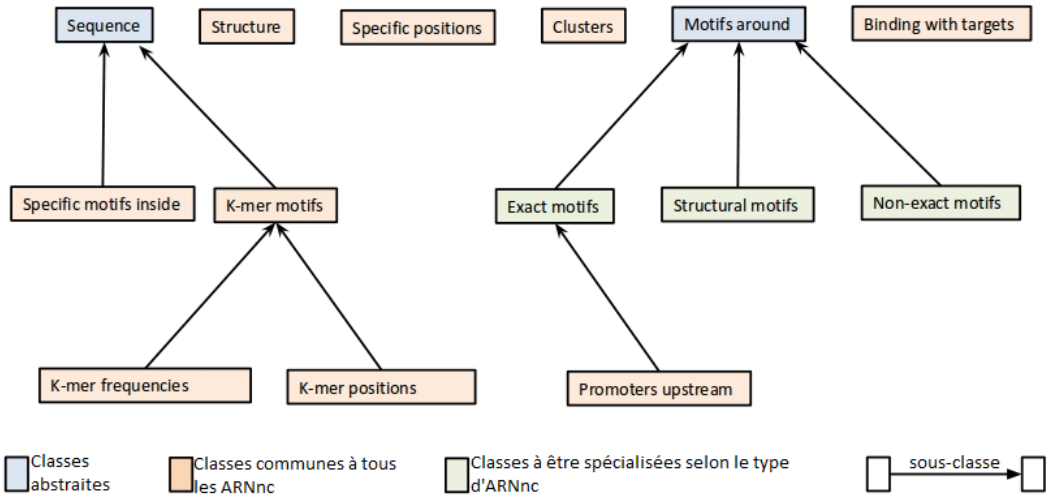


Figure 6.1: Les différentes classes de noyau définies dans IncRId et leur organisation hiérarchique

prédictives élevées, nous proposons d'utiliser un caractère non identifié $'X'$ dans l'alphabet des k-mers, qui peut être $'A'$, $'T'$, $'C'$ ou $'G'$, avec une probabilité d'occurrence maximale de 0,4.
En conséquence, nous générons 3 588 modèles qui représentent tous les k-mers possibles, pour $k = 1..5$. Après cela, nous effectuons une sélection supervisée sur cet ensemble. Cela représente un problème d'optimisation combinatoire où nous essayons d'identifier le sous-ensemble de k-mers le plus informatif pouvant atteindre une bonne prédiction. Cela peut être formulé en tant que problème de sélection de caractéristiques, où chaque k-mer est une caractéristique et chaque séquence se réfère à un échantillon. Pour effectuer la sélection, nous utilisons la méthode de sélection de caractéristiques CPM-FS proposée dans une première partie de cette thèse (Boucheham *et al.*, 2015a,b) qui effectue une sélection avec un nombre prédéfini de caractéristiques à sélectionner, afin d'identifier les N k-mers les plus représentatifs parmi tous ceux qui sont générés. Enfin, chaque séquence est représentée par un vecteur $N - dimensionnel$.

Fréquences des K-mers « K-mer frequencies »

Dans cette sous-classe de la classe "K-mer motifs", l'information discriminante utilisée pour effectuer la sélection des k-mers est leurs fréquences dans la séquence. Par la suite, le vecteur à N dimensions qui contient les fréquences des N k-mers divisés par la longueur de la séquence (pour raison de normal-

isation en fonction de la taille de la séquence).

Positions des K-mers « Kmer positions »

Dans cette deuxième sous-classe de la classe "K-mer motifs", l'information discriminative considérée est la position de chaque k-mer dans la séquence. Si un k-mer est présent plusieurs fois, nous gardons la position la plus proche du début de la séquence et s'il n'est jamais présent la valeur correspondante est zéro.

Clusters

Nous proposons ici de mettre en œuvre dans une approche d'apprentissage automatique supervisé la caractéristique d'apparition en clusters de certains types d'ARNnc. Á cet effet, on évalue le voisinage de chaque séquence, en calculant une matrice de distances entre la séquence considérée et ses K voisins sur le chromosome. Si la séquence a plusieurs positions sur le génome, nous choisissons la matrice avec la moyenne la plus faible, car elle représente une densité accrue et a plus de chances d'être près d'une formation de cluster. Enfin, chaque séquence est donc représentée par une matrice $(K+1)*(K+1)$.

Structure

La plus part des ARNnc ont une structure secondaire spécifique qui peut être utilisée comme une caractéristique discriminante. Afin de considérer cette caractéristique d'une manière computationnelle dans leur prédiction, une première étape est de toute évidence la prédiction de la structure secondaire. L'information de repliement ("Folding") de l'ARN est souvent représentée comme suit: un crochet d'ouverture ("(") pour indiquer les nucléotides appariés et un point (".") pour indiquer les nucléotides non appariés. Afin d'exploiter le mieux cette information, l'algorithme proposé consiste à s'intéresser à la structure de tous les triplets de nucléotides et à en mesurer leurs fréquences. Ces structures de triplets sont constituées par la combinaison du nucléotidique du milieu ($'A'$, $'T'$, $'C'$, ou $'G'$) de tous trois nucléotides adjacents, étant donné qu'il y a $8(2^3)$ compositions possibles de structure pour tous trois nucléotides adjacents, pour former $32(4x8)$ éléments de triplets différents qui contiennent à la fois l'information de repliement et de la séquence d'ARN. Ensuite, on compte les fréquences de chaque élément de triplet pour construire un vecteur 32-dimensionel représentera l'information de repliement de chaque séquence.

Liaison avec des cibles « Binding with targets »

Les principales informations mesurées et prises en compte dans cette classe de noyaux est le degré de la liaison entre les séquences d'ARNnc et des cibles spécifiques. Pour mesurer les informations de liaison séquence d'ARN / cible, plusieurs outils peuvent être utilisés, comme RNAPLEX (Tafer and Hofacker, 2008). Sur la base de l'information de repliement représentée par des crochets et des points, on utilise les structures de triplets pour chaque séquence, de la même manière que dans la classe « Structure ».

Positions spécifiques autour de la séquence « Specific positions »

Cette classe de kernels tient compte de la présence possible de l'ARNnc près ou à proximité de certaines observations spécifiques sur le génome. Pour intégrer cette information d'une manière computationnelle, nous mesurons la distance entre la séquence et ces observations. Comme chaque observation peut avoir une ou plusieurs positions sur le génome, nous avons besoin d'établir des critères de sélection selon le sens biologique de ces observations afin de choisir la meilleure position à considérer (généralement la plus proche). En outre, une séquence d'ARNnc peut avoir de nombreuses positions, donc la position qui a la distance la plus faible est conservée. Enfin, nous construisons un vecteur à N dimensions contenant les meilleures distances représentant les N observations.

Motifs autour de la séquence « Motifs around »

Développer une méthodologie d'implémentation générique pour tout élément appartenant à cette classe est une tâche difficile, car l'information discriminante à étudier dépend des spécificités biologiques de la caractéristique à considérer. Cependant, cette classe de noyaux est basée sur le génome de référence, afin de chercher des motifs en amont et / ou en aval de la séquence d'ARNnc. Généralement, nous estimons la distance la plus proche au motif et la longueur du motif découvert en tant qu'informations discriminantes ainsi que d'autres données spécifiques.
Plusieurs sous-classes seront donc construites selon le type de motifs recherchés, ce qui dépend du type d'ARNnc. On distingue particulièrement la classe de "Promoters upstream" (décrit ci-dessous) qui est commune à de nombreux types d'ARNnc.

Promoteurs en amont de la séquence « Promoters upstream »

Pour examiner le rôle d'un facteur de transcription dans la prédiction d'ARNnc, nous faisons usage des motifs de liaison identifiés et qui sont liés à leurs promoteurs. Dans la plupart des cas, il n'y a pas un motif de liaison explicite mais plutôt plusieurs motifs qui peuvent partager un motif consensus. Par conséquent, nous utilisons le génome de référence afin de le parcourir en amont de la position de chaque occurrence d'une séquence donnée d'ARNnc sur le génome. Nous partons du coté 5' de la séquence et nous recherchons en amont pour tous les motifs possibles et nous nous arrêtons au premier motif trouvé. Ainsi, nous maintenons trois types d'informations sur le motif découvert: la longueur du motif (L), la distance (D) entre le motif et la séquence, et une probabilité calculée comme suit: $4^L/D$. Cette probabilité permet la sélection de la position de l'ARNnc qui a le motif le plus proche à la séquence ainsi que le plus long au même temps.

6.2.4 Cas d'étude : la prédiction des ARNpi

Dans cette section, nous présentons un cas d'étude de notre méthodologie générique IncRId pour la prédiction des ARNnc: la prédiction des ARNpi.

Les caractéristiques des ARNpi dans divers organismes

Sur la base de l'approche intégrative proposée pour la caractérisation des ARNnc, nous avons examiné les études récentes de la littérature sur la biogenèse et la fonction des ARNpi, ainsi que sur d'autres observations biologiques liées à cette molécule dans diverses espèces, afin d'en déduire des caractéristiques intéressantes. Dans ce qui suit, nous résumons brièvement et classons ces caractéristiques qui sont principalement liés à: la fonction, la transcription et d'autres caractéristiques observées, comme le montre la Figure 6.2.

Caractéristiques liées à la fonction Des études récentes montrent que les deux première (5′ nucleotide) et dixième bases des ARNpi représentent une zone de liaison importante pour de nombreuses protéines Argonaute (Wang *et al.*, 2014b). En conséquence, les protéines Piwi et AUB montrent une forte préférence pour un uridine dans le coté 5′. Tandis que les proteines Ago3 associées aux ARNpi ne semblent pas tout enrichissement pour un 'U', mais ont tendance à contenir une adénosine dans leur dixième nucléotide, appelé aussi signature de ping-pong (Le Thomas *et al.*, 2014; Thomson and Lin, 2009).

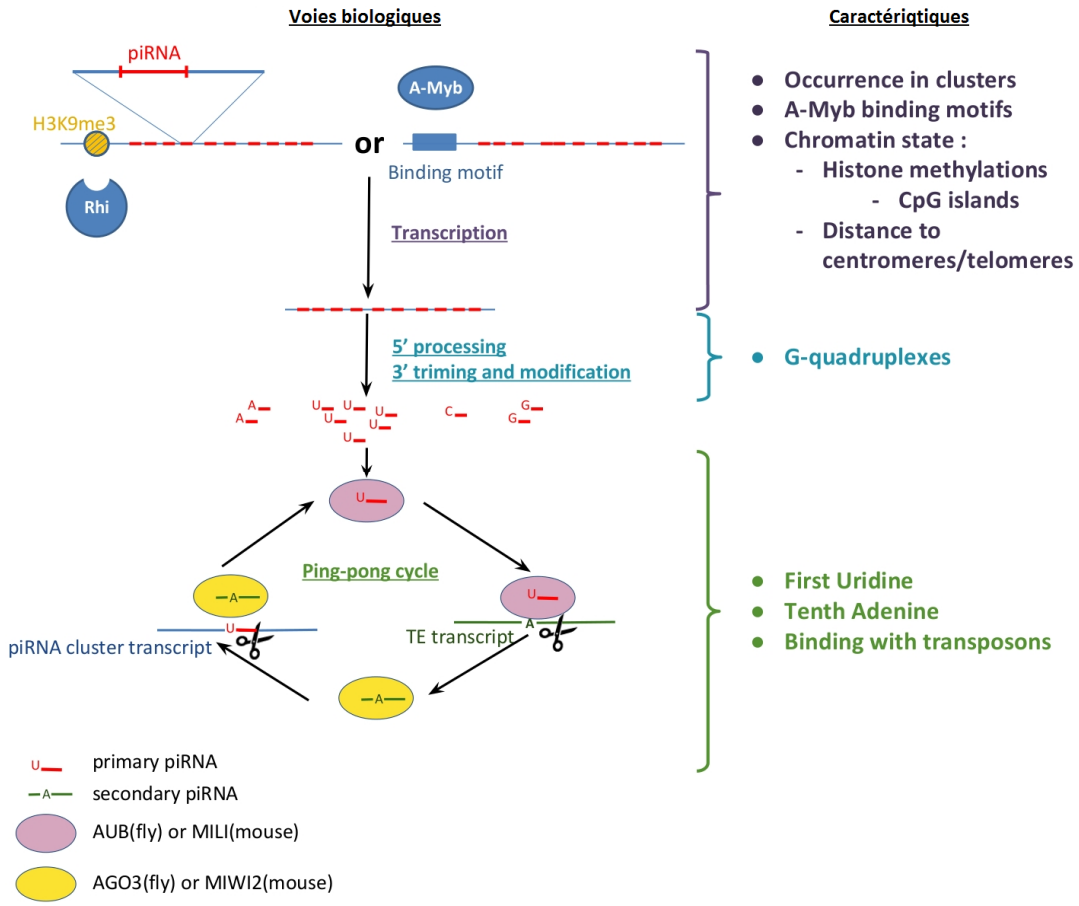


Figure 6.2: La relation entre la biogenèse des ARNpi (transcription, developement et fonction) et les caractéristiques mesurées: (i) les clusters d'ARNpi peuvent être transcrits si un histone particulier est méthylé (mouche) ou le promoteur A-Myb est à proximité (souris); (ii) G-quadruplex pourraient avoir un rôle dans le developement d'ARNpi et (iii) les deux premiere et dixième bases d'ARNpi (respectivement U et A) représentent une zone de liaison importante pour les protéines Argonaute, en participant au cycle "ping-pong" où les séquences d'ARNpi se lient à des transposons

Une autre caractéristique importante des ARNpi représente le rôle principal de ce type de petits ARN, qui est leur participation à l'inhibition de l'action des éléments transposables. Les ARNpi ont été trouvés à être antisens aux transposons et protègent le génome à partir d'éléments transposables invasives (TES) et conserve son intégrité (Weick and Miska, 2014).

Caractéristiques liées à la transcription Les ARNpi ont été montrés à apparaître en clusters chez les mammifères et les insectes (Thomson and Lin, 2009). Afin de mieux comprendre le processus de transcription, un point important est de prendre en compte l'état de la chromatine autour des séquences et d'envisager la quasi-totalité des modifications épigénétiques. Une étude récente indique que la plupart des clusters d'ARNpi dans Drosophila melanogaster ont été identifiés dans les régions hétérochromatines péricentromérique et télomérique (Yamanaka *et al.*, 2014). Par ailleurs, une autre étude sur les mêmes espèces rapporte que les clusters d'ARNpi sont souvent recouverts avec H3 triméthylée sur leur lysine 9 ($H3K9me3$). En outre, la transcription de certains clusters d'ARNpi nécessite "Rhino" qui est une protéine hétérochromatine 1 (HP1) homologue et a un chromodomaine (DR) qui se lie à H3K9me3 ou $H3K27me3$ (Yu *et al.*, 2015). Une autre façon d'envisager les modifications épigénétiques est de se basé sur la séquence génomique en prédisant des îles CpG (CpG islands) qui ont été montrées à être liées à la méthylation des histones (Rose and Klose, 2014). En outre, d'autres études ont étudié la transcription de clusters d'ARNpi dans "Mus musculus" et ont constaté que le facteur de transcription "AMyb" est nécessaire pour l'expression des ARNpi pachytènes. Il a été observé que A-Myb se lie à l'ADN à proximité du site de début de transcription des clusters d'ARNpi pachytènes (Li *et al.*, 2013).

D'autres caractéristiques observées Plusieurs clusters d'ARNpi ont été étudiés dans "Mus musculus", et certains sont encadrés par des répétitions inversées ("inverted repeats"), ce qui permet la formation de précurseurs contenant l'ARN double brin (Betel *et al.*, 2007). Dans la même étude, il a été également constaté que certains clusters d'ARNpi sont flanqués par des éléments transposables (TEs) tels que SINE, LINE et LTR. Cela a été rapporté aussi dans (Hirano *et al.*, 2014) ou il a été montré que la transposition de ces éléments peut être également dans les clusters meme. Enfin, une étude récente rapporte la présence de motifs "G-quadruplex" dans les clusters d'ARNpi des mammifères et ces structures peuvent avoir un rôle dans le développement des ARNpi (Vourekas *et al.*, 2015).

Le Tableau 6.1 résume les caractéristiques biologiques des ARNpi étudiées, avec pour chacune, les espèces dont lesquelles elle a été observée et/ou validée.

IpiRId: Outil de prédiction des ARNpi

IpiRId est une instanciation de notre outil générique IncRId. Il comprend de toute évidence les différentes classes de noyaux génériques décrites ci-

Caractéristiques	Espèce	Références
Première uridine	Fly, Mouse, Human, Rat, Nematode (C. elegans), Zebrafish and Silkworm (Bombyx mori)	(Wang *et al.*, 2014b; Weick and Miska, 2014) (Thomson and Lin, 2009)
Dixième Adénine	Human, Fly, Mouse, Zebrafish and Silkworm (Bombyx mori)	(Wang *et al.*, 2014b; Thomson and Lin, 2009)
Occurrence dans des clusters	Mammals and Insects	(Thomson and Lin, 2009)
Inhibition des transposons	Mammals and Insects	(Weick and Miska, 2014)
Les îles CpG	Mammals	(Rose and Klose, 2014; Kim and Kim, 2012)
G-Quadruplex	Human, Mouse, Rat and Macaque	(Vourekas *et al.*, 2015)
La présence des éléments transposables	Mouse and Marmoset	(Betel *et al.*, 2007; Hirano *et al.*, 2014)
Le promoteur A-Myb	Mouse	(Li *et al.*, 2013)
Répétitions inversées	Mouse	(Betel *et al.*, 2007)
Distance aux centromeres / télomères	Fly	(Le Thomas *et al.*, 2014)
la méthylation des histones	Fly	(Yu *et al.*, 2015)

Table 6.1: Les caractéristiques biologiques des ARNpi à travers les espèces

dessus, ainsi que d'autres classes de noyau spécifiques au ARNpi, qui sont spécialisées à partir des classes précédentes. Ces nouvelles classes formalisent les observations effectuées sur cet ARN, et qui concernent toutes la présence de certains motifs autour de la séquence. Ils sont donc une sous-classe de "Motifs around " classe (voir Figure 6.3). Évidemment, ces nouvelles classes, décrites ci-après, peuvent être utilisées/testées pour d'autres ARNnc.

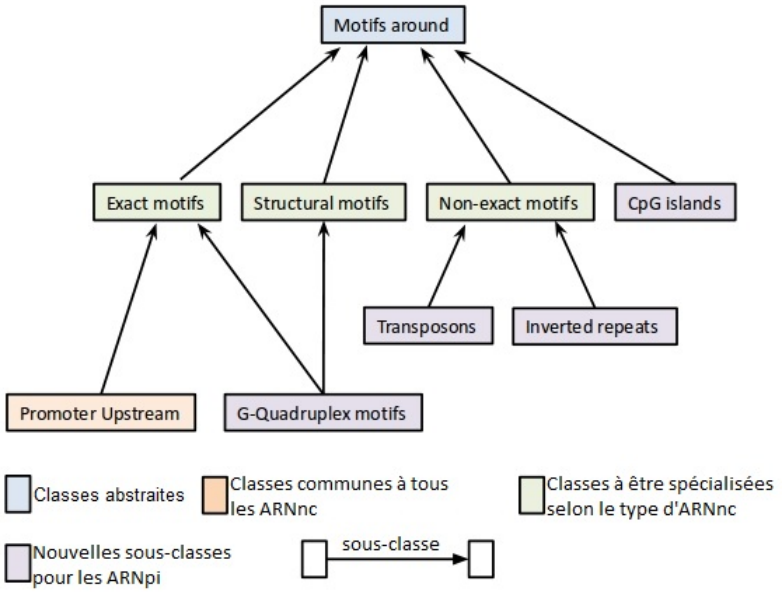


Figure 6.3: Les nouvelles classes de noyaux d'IpiRId qui sont des sous-classes de la classe générique "Motifs around"

Transposons Pour étudier la présence des TEs autour des ARNpi, le logiciel RepeatMasker [2] est utilisé. Nous nous attendons à trouver des TEs autour et dans les clusters d'ARNpi. À cette fin, nous regardons jusqu'à D kb en amont et D kb en aval des positions de chaque séquence d'ARNpi. Sur la base des sorties de ReapetMasker, nous calculons deux types d'informations: l'identité cumulée et la longueur cumulée pour chaque élément transposable. L'identité est calculée comme suit:

$$Identit = 1 - RM - RD - RS \tag{6.5}$$

où RM, RD et RS sont, respectivement, le ratio de mésappariement, de suppression et de substitution. Si la séquence d'ARNnc a plusieurs positions sur le génome, nous choisissons la position avec l'identité cumulée la plus élevée. Du fait que les différents TEs n'ont pas la même chance d'être trouvés autour des séquences d'ARNpi (par exemple, des éléments LINE sont en effet plus souvent trouvés dans les clusters d'ARNpi que les éléments SINE puis les LTR (Hirano *et al.*, 2014)), ils sont donc pondérés convenablement. Enfin, chaque séquence est représentée par un vecteur à $(2 * N)$ dimensions, où N est le nombre d'éléments transposables (TE) considérés.

Les répétitions inversées "Inverted repeats" Afin d'évaluer la présence des répétitions inversées à proximité d'une séquence donnée, on utilise la méthode de détection utilisée par (Betel *et al.*, 2007). Nous faisons usage de la séquence génomique D kb en amont et en aval de la séquence et nous comparons la séquence obtenue à son complément avec BLAST (bl2seq) (Tatusova and Madden, 1999). Les alignements de plus de 20 bases et avec plus de 90% d'identité sont considérées. En conséquence, nous calculons la moyenne de leurs longueurs et le nombre cumulé de leurs identités. Chaque séquence est donc représentée par un vecteur à deux dimensions.

G-quadruplex Nous cherchons ici des structures G-quadruplex dans le voisinage de chaque séquence. Pour ce faire, nous utilisons un script Python [3] permettant de prédire les G-quadruplex D kb en amont et en aval sur le brin de la séquence ainsi que sur le brin opposé. Ensuite, nous calculons cinq types d'informations: la distance au plus proche G-quadruplex sur le brin de la séquence, la distance au plus proche G-quadruplex sur le brin opposé, le nombre d'occurrences des G-quadruplex sur les deux brins, et enfin la longueur cumulée de tous les G-quadruplex. Chaque séquence est donc

[2]Le logiciel RepeatMasker est disponible à: `http://www.repeatmasker.org`

[3]quadparser.py, disponible à `http://bioinformaticsmisc.googlecode.com/svn-history/r16/trunk/quadparser.py`

Kernel	Classe	Paramètres d'instanciation
U1|A10	Specific motifs inside	{motif,position}: {U,1}, {A,10}
K-merFreq	K-mer frequencies	N (number of k-mers) = 32 motifs
K-merPos	K-mer positions	N (number of k-mers) = 32 motifs
TE binding	Binding with targets	target: Transposable elements (TE)
CentroTelo	Specific positions	observation: centromer, telomeres
Histone	Specific positions	observation: H3K9me3, H3K27me3
Cluster	Clusters	K (number of neighbours) = 4
A-Myb	Promoters upstream	promoter: A-Myb; D = 40 kb
G-Quadruplex	G-quadruplex	D = 40 kb
CpG islands	CpG islands	L = 100 ; D = 20 kb
LINE|SINE|LTR	Transposons	TE: LINE, SINE, LTR; D = 40 kb
InvertRep	Inverted repeats	D = 40 kb

Table 6.2: l'instanciation des noyaux d'IpiRId. (D: distance; L: longueur minimale)

représentée par un vecteur de 5 dimensions. Si une séquence a de multiples positions, nous choisissons la position avec le plus proche G-quadruplex sur son brin.

Les îles CpG "CpG islands" Nous considérons également la méthylation différemment en utilisant uniquement la séquence génomique en amont d'un ARNpi donné et prévoir les îlots CpG sur cette séquence. A cet effet, on utilise l'outil newcpgreport (Rice *et al.*, 2000), pour détecter les îlots CpG avec une longueur minimale de L nucléotides. Pour chaque séquence, la séquence génomique D Kb en amont est donnée à newcpgreport qui calcule les informations connexes: distance à l'île CpG la plus proche, le nombre des îles CpG prédites, la moyenne du ratio exprimé observé, la moyenne des longueurs des îles et la moyenne de la somme des bases $C + G$ dans les îles. Chaque séquence est alors représentée par un vecteur de 5 dimensions. Si une séquence a de multiples positions, nous choisissons la position avec la distance la plus faible à une île CpG prédite en amont.

Pour résumer, IpiRId est actuellement composé de douze noyaux (kernels) qui sont énumérés dans le Tableau 6.2.

6.3 Résulats et discussions

6.3.1 Construction des jeux de données

Dans un cadre de classification binaire, le jeu de données d'apprentissage doit être constitué d'éléments représentatifs des classes à évaluer. Ici nous nous intéresserons à chaque fois à deux classes : la classe d'ARN à prédire et les ARNs qui ne font pas partie de cette classe.
Afin de créer nos jeux de données d'apprentissage, nous construisons trois ensembles de données avec des séquences positives et négatives d'ARNpi dont chacun se réfère à l'une des trois espèces considérées dans cette étude: humain (Homo sapiens), souris (Mus musculus) et la mouche (Drosophila melanogaster). Les séquences povitives et non redondantes d'ARNpi ont été récupérées à la fois des deux bases de données piRNAbank (Lakshmi and Agrawal, 2008) (`http://pirnabank.ibab.ac.in/`) et piRBase (Zhang *et al.*, 2014) (`www.regulatoryrna.org/database/piRNA/`), d'où nous avons téléchargé 32208 (humain), 39986 (souris) et 18508 (mouche) séquences d'ARNpi. Comme pour les séquences négatives (non ARNpi), nous considérons différents types d'ARNnc:

- 449, 244 et 93 séquences d'ARNt de l'humain, la souris et la mouche, respectivement, téléchargées à partir de la base de données génomique d'ARNt (`http://lowelab.ucsc.edu/GtRNAdb/`).
- 1747, 712 et 288 séquences de micro-ARN matures de haute confiance ("high-confidence") (Kozomara and Griffiths-Jones, 2013) de l'humain, la souris et la mouche, respectivement, téléchargées à partir de miRBase (`http://www.mirbase.org/`).
- 9113, 4896 et 740 séquences de régions exoniques de l'humain, la souris et la mouche, respectivement, de longueur entre 25 à 33 pour l'humain et la souris et de 22 à 35 pour mouche, téléchargées à partir Ensembl (`www.ensembl.org/index.html`).

Toutes les séquences positives et négatives ont été alignées sur les génomes de références : hg38 humain, mm10 souris et dm6 mouche on utilisant le logiciel Bowtie (Langmead et al., 2009), qui est utilisé aussi par piRBase pour déterminer les positions génomiques (Zhang *et al.*, 2014), sans permettre de gaps et avec un maximum d'un mésappariement (mismatch=1) pour les séquences qui ne match pas exactement. Sauf pour les micro-ARNs qui sont inclus dans des précurseurs (pré-ARNmi), leur réalignement produit trop de positions. Ainsi, nous utilisons les positions données par miRBase et les

soulever au génome de référence approprié en utilisant l'outil "liftover" depuis "UCSC Genome Browser" (Hinrichs *et al.*, 2006).
En outre, les transposons ont été récupérés à partir de la table "rmsk" de l'UCSC Genome Browser (Hinrichs *et al.*, 2006) à l'exclusion de ceux avec une annotation "rich", nucléotides répétés et transposons redondants. Pour des raisons de calcul, nous considérons seulement les transposons avec une longueur comprise entre 35 et 100 nt, et finalement nous choisissons aléatoirement 1000 transposons de tout l'ensemble. Cela nous permet de regarder le même nombre de transposons de longueur similaire pour chaque espèce et de mieux comparer la pertinence de cette caractéristique entre les espèces. Cette longueur est un paramètre qui peut être fixé par l'utilisateur.
D'autre part, les données épigénétique "Chip-Seq" représentées par les positions des histones H3K9me3 et H3K27me3 ont été prises depuis le dépôt épigénétique de l'NCBI. Les tissus / cellules considérées dans notre étude sont: les cellules T pour l'humain ([4]), les cellules souches embryonnaires pour la souris ([5]) et des ovaires (pour H3K9me3) et les testicules (pour H3K27me3) pour la mouche ([6]). Nous utilisons également l'outil "liftover" (Hinrichs *et al.*, 2006) pour soulever des assemblages de données épigénétiques téléchargées à ceux appropriés adoptés pour chaque espèce considérée.

Le Tableau 6.3 résume les différents ensembles de données téléchargés utilisées dans notre étude.

Espèces/Données	povitive	négative			chip-seq		transposons	génome de référence
	ARNpi	ARNt	ARNmi	régions exoniques	H3K9me3	H3K27me3		assemblage
Homo sapiens	32208	449	1747	9113	6346007	8968536	903140	hg38
Mus Musculus	39986	244	712	4896	2751	1232402	3504253	mm10
Drosophila melanogaster	18508	93	288	740	508	2322	803255	dm6

Table 6.3: Les données téléchargées à travers les espèces: nombre de séquences dans les ensembles de données positives et négatives utilisées dans nos expérimentations et les différentes autres sources de données utilisées par notre approche intégrative: nombre de positions utilisées pour les histones méthylés ainsi que des transposons

[4]téléchargé à partir: `http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx`

[5]téléchargé à partir: `http://www.ncbi.nlm.nih.gov/epigenomics/166` for H3K9me3 et `http://www.ncbi.nlm.nih.gov/epigenomics/164` for H3K27me3

[6]téléchargé à partir: `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1121659` pour H3K9me3 et `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM480447` pour H3K27me3

Mesures de prédiction des résultats

Les résultats ci-dessous sont donnés en fonction de cinq mesures, généralement utilisées dans la classification supervisée. Les formules de ces mesures sont données ci-après en utilisant les abréviations suivantes: vrai positif (TP), faux positifs (FP), vrai négatif (TN) et faux négatifs (FN).

$$Sensitivity(Se) = \frac{TP}{TP + FN} * 100 \tag{6.6}$$

$$Specificity(Sp) = \frac{TN}{TN + FP} * 100 \tag{6.7}$$

$$Precision(Pre) = \frac{TP}{TP + FP} * 100 \tag{6.8}$$

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{6.9}$$

$$F1\ score(F1) = \frac{2TP}{2TP + FP + FN} * 100 \tag{6.10}$$

6.3.2 Comparaison des outils de prédiction des ARNpi

Peu d'outils ont été proposés récemment pour prédire les ARNpi. Le premier outil publié est piRNApredictor qui est basé sur les fréquences de certains motifs (k-mer) (Zhang *et al.*, 2011b). Il utilise la méthode Fisher pour sélectionner les k-mers les plus discriminants ($k = 1...5$), puis effectue une amélioration de Fisher avec un seuil pour classer les séquences. Un autre outil basé k-mer, appelé Pibomd, a été récemment proposé (Liu *et al.*, 2014). Il recherche pour tous les 5-mer et 4-mer les motifs avec 3 nucléotides communs appartenant à 40% des séquences d'apprentissage. Les fréquences de tous les k-mers trouvés sont ensuite utilisées dans un SVM pour classer les séquences de prédiction. Un autre outil proposé récemment, appelé Piano, est basé sur l'information de liaison ARNpi/transposon (Wang *et al.*, 2014a). Il utilise l'outil "SeqMap" pour sélectionner les séquences avec maximum trois mismatches et l'outil "RNAplex" pour faire le repliement de chaque séquence avec des transposons, avec en fin un SVM pour faire de la prédiction.
Pour procéder à une comparaison entre notre outil et les autres outils existants, nous avons réentrainé ces outils sur nos ensembles de données en utilisant la technique de validation croisée (5-cross validation). Il faut noter que nous avons eu beaucoup de problèmes pour réentrainer ces outils, car ils sont fonctionnels uniquement en mode prédiction, et l'information pour le réentrénement n'est pas mentionnée dans leurs manuels ou publications.

Comme décrit ci-dessus, trois espèces ont été prises en compte dans notre étude: humain, souris et mouche. Nous avons construit un ensemble de données contenant 5000 séquences ARNpi et 5000 séquences pseudo ARNpi pour les deux espèces humain et souris et 1100 séquences ARNpi et 1100 séquences pseudo ARNpi pour mouche. Ces séquences ont été obtenues par une sélection aléatoire à partir des jeux de données initiales téléchargés pour chaque espèce.

Outil/Espèce	Humain					Souris					Mouche				
	Acc	Se	Sp	Pre	F1	Acc	Se	Sp	Pre	F1	Acc	Se	Sp	Pre	F1
piRNApredictor	71.85+-1.53	48.40	**95.5**	**91.49**	63.30	70.95+-1.15	47.79	94.10	89.01	62.19	52.17+-3.72	63.90	40.45	51.76	57.19
Piano	50	0	100	0	0	50	0	100	0	0	87.9+-1.472	78.90	96.90	96.22	86.70
Pibomd	78.13+-1.38	78.05	78.21	78.17	78.11	79.13+-1.19	79.43	78.82	78.94	79.18	66.08+-4.02	70.44	61.72	64.78	67.94
piRPred	81.20+-1.25	80.54	81.86	81.67	81.07	90.92 +-0.51	90.36	91.48	91.39	90.87	86.36+-2.33	86	86.72	86.66	86.30
IpiRId	**90.09+-0.25**	**90.56**	89.62	89.73	**90.13**	**93.66+-0.46**	**90.74**	**96.58**	**96.37**	**93.47**	**92.59+-1.87**	**87.27**	**97.90**	**97.67**	**92.12**

Table 6.4: Comparaison des performances: résultats de 5-fold cross-validation d'IpiRId et d'autres outils existants en fonction de: accuracy (ACC), sensibilité (Se), spécificité (Sp), précision (Pre) et le F1 score (F1)

Le Tableau 6.4 présente les résultats de validation croisée de notre outil IpiRId et les autres outils existants (piRNApredictor, Piano, Pibomd et piRPred) à travers les trois espèces. Les résultats sont donnés en fonction de cinq mesures, généralement utilisées dans les tâches de classification supervisée: la sensibilité (Se), spécificité (Sp), Précision (PRE), Accuracy (ACC) et le F1 score (F1). Ils montrent clairement la surperformance de notre outil. IpiRId donne plus de 90% de précision dans toutes les espèces, ainsi que des valeurs proches de sensibilité, spécificité, précision et F1 score qui sont tous autour de 90%.
Pibomd, l'outil montrant les deuxièmes meilleurs résultats, donne une précision, ainsi qu'une sensibilité, spécificité et F1 score moins de 80% dans toutes les espèces (moins de 70% dans la mouche). Notez que Piano ne fonctionne que sur Drosophila melanogaster. Cela pourrait être dû au fait que plus de 70% d'ARNpi peuvent être alignés sur des transposons pour cette espèce, en utilisant SeqMap, ce qui n'est pas le cas pour l'humain et la souris. Nous pensons que, pour cette raison, dans (Wang *et al.*, 2014a), les auteurs n'ont pas montré des résultats de validation croisée de l'humain et la souris, mais plutôt ils ont utilisé le modèle entrainé sur la mouche afin de faire la prédiction sur ces espèces.
Les espaces ROC donnés dans la Figure 6.4 et correspondant aux résultats de validation croisée obtenus par IpiRId, piRPred, Pibomd, Piano et piRNApredictor montrent clairement que IpiRId donne le meilleur compromis entre la spécificité et la sensibilité dans toutes les espèces considérées, en par-

ticulier pour la souris et la mouche . Les autres outils donnent des résultats très hétérogènes à travers les espèces.

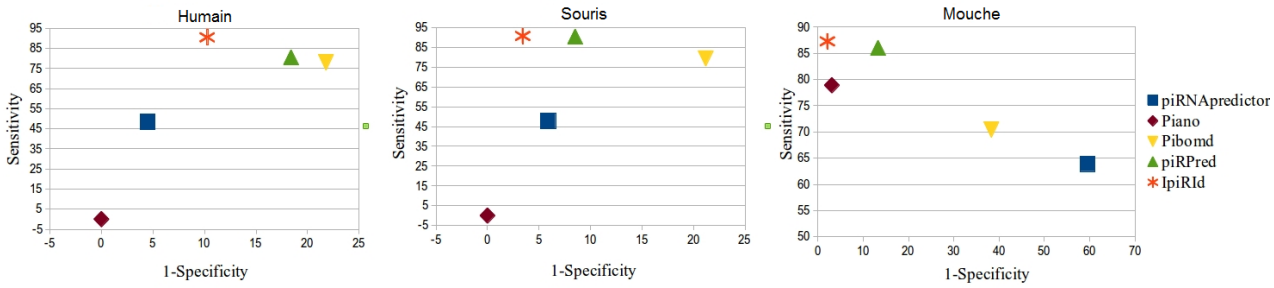


Figure 6.4: Les espaces ROC des résultats de validation croisée d'IpiRId et d'autres outils à travers les espèces.

6.3.3 Pertinence des caractéristiques à travers les espèces

Un intérêt significatif de notre outil est qu'il permet aux biologistes de mesurer la pertinence d'une caractéristique donnée dans les espèces considérées. Il est évident que les caractéristiques sont souvent observées expérimentalement en une ou plusieurs espèces, comme montré dans le Tableau 6.1. Ici, nous présentons les résultats de prédiction qui montrent la pertinence de chacune de ces catachétiques dans les trois espèces étudiées.
Les résultats présentés dans la Figure 6.5 confirment que les caractéristiques du premier "U", du dixième "A", la présence en clusters et la liaison avec des transposons, qui ont été observées dans plusieurs espèces, à savoir les mammifères et des insectes, sont ceux qui mieux caractérisent les ARNpi dans toutes les espèces étudiées. Les noyaux implémentant ces caractéristiques sont en effet ceux qui donnent les meilleurs résultats de prédiction (entre 70 et 91% de précision). Notez que les résultats obtenus par le noyau de liaison avec TE peuvent certainement être améliorés en tenant compte d'un ensemble plus large de transposons (pour des raisons de calcul, nous avons considéré un ensemble de seulement 1000 transposons dans cette étude).
En outre, les deux noyaux liés au k-mer (fréquences et positions de K-mer), qui ne sont pas spécifiques au ARNpi et pourraient être utilisés pour tous les ARNnc, donnent de bons résultats. Ce qui valide la nouvelle méthodologie proposée dans cette thèse pour l'identification des k-mers les plus discriminants. En ce qui concerne les autres noyaux représentant des caractéristiques observées chez des espèces spécifiques, les résultats sont différents d'une es-

pèce à une autre, ce qui suggère que ces caractéristiques ne sont pas conservées dans toutes les espèces. Par exemple, deux caractéristiques observées dans la mouche ne semblent pas conservées dans les autres espèces considérées: la distance aux centromères et télomères et la méthylation des histones. Les kernels qui implémentent chacune de ces deux caractéristiques donnent de très bons résultats de prédiction chez la drosophile, avec une précision d'environ 90%, tandis que, dans les espèces humain et souris, ils donnent une précision inférieure à 70%. La caractéristique G-quadruplex, observée chez l'humain et la souris (et aussi chez le rat et Macaque), donne des résultats de précision similaires, autour de 60%, sur les trois espèces testées, ce qui montre que cette caractéristique n'est pas très importante, même si elle ne semble pas être due à un événement aléatoire. En outre, nous pouvons faire la même remarque sur la caractéristique de la présence de transposons en amont ou en aval de la séquence d'ARNpi, qui a été observé chez la souris et l'Ouistiti (Marmoset). Le noyau correspondant (LINE | SINE | LTR) donne relativement les mêmes résultats de précision dans les différents espèces (environ 60%).

Figure 6.5: La pertinence des caractéristiques d'IpiRId à travers les espèces: souris, humain et mouche.

De manière surprenante, deux caractéristiques observées dans la souris ne donnent pas de résultats significatifs de prédiction dans cette espèce: le facteur de transcription (promoteur A-Myb) et les clusters d'ARNpi encapsulés par une répétition inversée. Les noyaux implémentant ces deux caractéristiques donnent respectivement environ 50% et 60% de précision. Les résultats sont assez semblables pour l'humain, cependant, ils donnent des résultats de précision relativement bons, autour de 74% dans la mouche. Notez que concernant le promoteur A-Myb, la faible précision pourrait être pour la raison que cette caractéristique caractérise une sous-classe particulière d'ARNpi, les ARNpi pachytènes.
Un autre résultat remarquable concerne le noyau îles CpG. Du fait que cette

caractéristique est liée à la méthylation des histones chez les mammifères, nous nous attendions à obtenir des résultats de prédiction proches de ceux obtenus par le noyau histone méthylation. Mais ce n'est pas le cas, puisque pour l'humain le noyau CpG donne une précision de 75%, tandis que le noyau d'histone méthylation donne seulement 58%. Inversement, dans la mouche, il donne une précision de 63% lorsque le noyau d'histone méthylation donne plus de 90% précision.
Pour résumer, on peut observer que les différentes espèces étudiées partagent très peu de caractéristiques. Cependant, notre méthode faire face à cette limitation et permet d'obtenir de bons résultats de prédiction en utilisant toutes ces caractéristiques ensemble.

6.4 Conclusions

Ce chapitre présente deux principales contributions dans le domaine de la prédiction d'ARNnc:

- Une approche globale et intégrative basée sur la méthodologie MKL (Multiple Kernel Leurning) en tenant compte d'un grand ensemble de caractéristiques hétérogènes, et traitant la non-conservation de certaines caractéristiques entre les espèces (prenant ainsi en compte l'évolution de l'espèce).On a fait une étude approfondie sur les caractéristiques biologiques possibles qui caractérisent les ARNnc et qui pourraient être utilisées pour leur prédiction par des méthodes computationnelles. Ensuite, nous avons classé ces caractéristiques en plusieurs classes principales et les implémentées dans des noyaux génériques, qui pourraient être utilisées soit directement ou adaptées à un type spécifique d'ARNnc. Ainsi, nous proposons un outil générique et modulaire qui pourrait être facilement utilisé pour toute catégorie d'ARNnc. Par conséquent, il permet de tester les caractéristiques observées dans un type d'ARNnc sur d'autres, ainsi que de tester la validité de nouvelles caractéristiques qui n'ont jamais été considérées.

- Un outil pour la prédiction des ARNpi appelé IpiRId, qui est une application/instanciation du cadre générique proposé pour la prédiction des ARNnc. Nous avons fait un état de l'art sur les ARNpi, afin d'en déduire des caractéristiques qui pourraient être utilisées pour leur prédiction. Le résultat de cette étude est l'extraction d'un grand nombre de caractéristiques hétérogènes (13 caractéristiques, dont très peu ont déjà été prises en compte dans des outils informatiques), principalement liées à la fonction et la transcription. IpiRId surpasse tous les outils

existants pour la prédiction d'ARNpi, donnant une précision d'environ 90% chez l'humain, la souris et la mouche. Enfin, et grâce à notre méthode MKL et l'outil modulaire, nous avons pu mesurer l'importance de chaque caractéristique dans ces trois espèces.

Notre outil est directement disponible en trois modes d'exécution:

- mode Prediction (pour les biologistes) en utilisant les modèles de prédiction fournis pour chaque espèce,
- mode Cross-validation (pour les chercheurs bio-informaticiens)
- et finalement le mode TrainPredict (faire l'entrainement sur un jeux de données et ensuite utiliser le modèle de prédiction de sortie pour faire la prédiction).

Ces trois modes peuvent être utilisés à partir du serveur Web convivial (user-friendly web server) ou la version autonome (standalone version) disponibles à: `https://evryrna.ibisc.univ-evry.fr/IpiRId/`.

Les travaux futurs concernent l'intégration d'autres noyaux implémentant des caractéristiques spécifiques à d'autres types d'ARNnc (micro-ARN (ARNmi), ARNsno, ARNcirc...) en effectuant des études approfondies sur leurs caractéristiques récemment découverts. En conséquence, nous allons instancier pour chaque type un outil spécifique de la même manière qu'avec les ARNpi.

Chapitre

7 Conclusion générale et perspectives

Dans cette thèse, nous avons abordé le problème de l'identification de biomarqueurs moléculaires dont le but est d'arriver à sélectionner des signatures moléculaires hautement informatives et discriminatives pour certains types du cancer. Nous nous sommes intéressés à la découverte de biomarqueurs au niveau génomique et transcriptomique, en raison de leur extrême importance dans la médecine personnalisée. La nature du problème à traiter change d'un niveau « omique » à un autre, en fonction des disponibilités de connaissances et de données biologiques obtenues principalement des nouvelles biotechnologies à haut débit. Ainsi, l'identification de biomarqueurs génomique, se base sur des données d'expression géniques extraites des expériences bio-puces et s'appuie principalement sur les algorithmes de sélection de caractéristiques supervisés. En revanche, la découverte de biomarqueurs transcriptomiques, revient à identifier des ARN non-codants à partir des données RNAseq ainsi les classer par la suite selon certaines caractéristiques spécifiques à chacun de leur type.
Dans un premier temps, Nous avons étudié les différents aspects concernant l'application des méthodes de sélection de caractéristique basées ensemble à l'identification de biomarqueurs. En effet, notre étude atteste que cette technique est une voie prometteuse pour avoir une sélection plus stable et précise dans l'identification des gènes du cancer. Nous avons également proposé une approche parallèle sur la base d'un méta-ensemble de filtres pour la découverte de biomarqueurs à partir des données de grande dimension. La méthode MPME-FS est différente des autres méthodes de sélection de caractéristiques d'ensemble car elle effectue une sélection parallèle en deux étapes: la première au sein de chaque ensemble par l'agrégation des résultats des différents sélecteurs, tandis que la deuxième étape porte sur l'agrégation des résultats de tous les ensembles en utilisant une deuxième fonction de consensus. En outre, MPME-FS est rapide comme elle n'utilise pas d'algorithme d'apprentissage dans le processus de sélection. Nous nous sommes intéressés par la suite aux méthodes wrapper qui permettent de gagner plus de précision de classification. Nous avons proposé une méthode hybride wrapper/fil-

ter de sélection de caractéristiques sur la base de la coopération parallèle de métaheuristiques (CPM-FS) pour la sélection d'un nombre prédéterminé de biomarqueurs utiles. La méthode proposée emploie différentes métaheuristiques basées population avec une nouvelle stratégie d'initialisation basée sur un ensemble de filtres. Ainsi, un mécanisme de réparation de solutions à base de filtre afin de réparer les individus manipulés dans le processus de sélection et contribuer également dans la bonne exploration de l'espace de recherche de solutions. Ensuite, nous avons proposé une méthode de sélection de gènes en deux étapes dont chacune est basée wrapper en utilisant la méthode précédemment proposée (CPM-FS) et une fonction de consensus qui permet de prendre en compte les dépendances entre les gènes.
Les méthodes citées au-dessus peuvent être facilement étendues à d'autres problèmes de sélection de caractéristiques et sont adaptées aux grands ensembles de données.
Dans un second temps, nous avons proposé une approche globale et intégrative basée sur une technique d'apprentissage à noyaux multiples (Multiple Kernel learning) en prenant en compte un grand nombre de caractéristiques hétérogènes, et en traitant la non-conservation de certaines caractéristiques entre les espèces (prenant ainsi en compte l'évolution des espèces). Nous avons réalisé une étude approfondie sur les caractéristiques biologiques possibles qui caractérisent les ARNnc et qui pourraient être utilisées pour leur prédiction par des méthodes computationnelles. Ensuite, nous avons classé ces caractéristiques en plusieurs classes principales et les implémentées dans des noyaux génériques, qui pourraient être utilisés soit directement ou adaptés à un type d'ARNnc d'intérêt. Ainsi, nous avons proposé un cadre générique et modulaire qui pourrait être facilement utilisé pour toute catégorie d'ARNnc. Il permet également de tester les caractéristiques observées pour un type d'ARNnc sur d'autres, ainsi que d'examiner la validité des nouvelles caractéristiques qui n'ont jamais été prises en considération.
Par la suite, nous avons proposé un outil pour la prédiction des ARNpi appelé IpiRId, qui est une application / instanciation du cadre générique proposé pour la prédiction des ARNnc. Nous avons fait un état de l'art sur les dernières connaissances sur les ARNpi, afin d'en déduire des caractéristiques qui pourraient être utilisées pour leur prédiction. Cette étude a abouti à un grand nombre de caractéristiques hétérogènes (13 caractéristiques), principalement liées à la fonction et la transcription. IpiRId implémente toutes ces caractéristiques, très peu d'entre elles ont déjà été prises en compte dans des outils bioinformatiques. IpiRId surpasse tous les outils existants pour la prédiction d'ARNpi, donnant une précision d'environ 90% chez l'homme, la souris et la mouche. Enfin, et grâce à notre méthode basé MKL et notre outil modulaire, nous avons pu mesurer l'importance de chaque caractéris-

tique dans ces trois espèces.

Ce travail étant encore préliminaire, nous avons identifié plusieurs perspectives à court terme d'évolution pour améliorer la prédiction des ARNnc, telle que l'intégration d'autres noyaux implémentant des caractéristiques spécifiques à d'autres types d'ARNnc (microARN, ARNsno, ARNcirc...) en effectuant des études approfondies sur leurs caractéristiques récemment découvertes. En conséquence, nous allons instancier pour chaque type d'eux un outil spécifique de la même manière qu'avec les ARNpi. Pour ce qui est approches proposées pour la sélection de caractéristiques, bien qu'ils nous aient permis d'obtenir des résultats très compétitifs, nos approches pourraient être encore améliorées. Par exemple, nous pourrions utiliser et tester d'autres filtres et d'autres métaheuristiques basées population dans la méthode hybride wrapper /filter (CPM-FS) afin d'améliorer l'exploration de l'espace de recherche de solutions et par conséquent avoir de meilleurs résultats de prédiction.
Nos travaux offrent également des perspectives à plus long terme. À titre d'exemple, nous envisageons intensifier notre approche intégrative pour la prédiction des ARNnc ainsi que les méthodes de sélection de caractéristiques dans le cadre MapReduce pour faire face aux très grands volumes de données (Big data). Nous comptons également appliquer et tester les méthodologies développées pour la recherche de biomarqueurs, gènes et ARNs (notamment les microARN), dans le cancer avec validation clinique.

Production Scientifique

Articles conference

1. **Anouar Boucheham** and Mohamed Batouche. "Robust biomarker discovery for cancer diagnosis based on meta-ensemble feature selection". **In : Science and Information Conference (SAI), 2014. IEEE, 2014. p. 452-560.**

2. **Anouar Boucheham** and Mohamed Batouche. "Découverte de biomarqueurs pour le diagnostic du cancer par des approches basées sur l'intelligence computationnelle". **Doctoriales de l'Université De Boumerdès 09 au 14 Mars 2014.**

3. **Anouar Boucheham** and Mohamed Batouche. "Breast Cancer Classification Approach based on Biomarker Discovery and Ensemble of Classifiers". **In: CISC 2014, 29.**

4. **Anouar Boucheham**, Mohamed Batouche, and Souham Meshoul. "An Ensemble of Cooperative Parallel Metaheuristics for Gene Selection in Cancer Classification". **In : Bioinformatics and Biomedical Engineering. Springer International Publishing, 2015. p. 301-312.**

Articles journal

1. **Anouar Boucheham** and Mohamed Batouche. "Massively Parallel Feature Selection Based on Ensemble of Filters and Multiple Robust Consensus Functions for Cancer Gene Identification". **Intelligent Systems in Science and Information 2014. Springer International Publishing, 2015. p. 93-108.**

2. **Anouar Boucheham**, Mohamed Batouche and Souham Meshoul. "Robust hybrid wrapper/filter biomarker discovery from gene expression data based on generalised island model". **Int. J. of Computational Biology and Drug Design, vol. 8, no. 3, pp. 251-274, 2015.**

Bibliographie

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, **26**(3), 392–398.

Alba, E., García-Nieto, J., Jourdan, L., and Talbi, E.-G. (2007). Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, pages 284–290. IEEE.

Almuallim, H. and Dietterich, T. G. (1994). Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, **69**(1-2), 279–305.

Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., and Napolitano, A. (2012). A review of the stability of feature selection techniques for bioinformatics data. In *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*, pages 356–363. IEEE.

Azuaje, F. (2011). *Bioinformatics and biomarker discovery:" omic" data analysis for personalized medicine*. John Wiley & Sons.

Baldi, P. and Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**(6), 509–519.

Bao, L. and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, **21**(10), 2185–2190.

Ben-Bassat, M. (1982). Pattern recognition and reduction of dimensionality. *Handbook of Statistics*, **2**, 773–910.

Bermejo, P., Gámez, J. A., and Puerta, J. M. (2011). A grasp algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets. *Pattern Recognition Letters*, **32**(5), 701–711.

Bernal, A., Crammer, K., Hatzigeorgiou, A., and Pereira, F. (2007). Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput Biol*, **3**(3), e54.

Betel, D., Sheridan, R., Marks, D. S., *et al.* (2007). Computational analysis of mouse piRNA sequence and biogenesis. *PLoS Comput Biol*, **3**(11).

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Blanco, R., Larrañaga, P., Inza, I., and Sierra, B. (2004). Gene selection for cancer classification using wrapper approaches. *International Journal of Pattern Recognition and Artificial Intelligence*, **18**(08), 1373–1390.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, **97**(1), 245–271.

Boln-Canedo, V., Snchez-Maroo, N., and Alonso-Betanzos, A. (2016). *Feature Selection for High-Dimensional Data*. Springer.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2014a). Data classification using an ensemble of filters. *Neurocomputing*, **135**, 13–20.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J., and Herrera, F. (2014b). A review of microarray datasets and applied feature selection methods. *Information Sciences*, **282**, 111–135.

Boucheham, A., Batouche, M., and Meshoul, S. (2015a). An ensemble of cooperative parallel metaheuristics for gene selection in cancer classification. In *Bioinformatics and Biomedical Engineering*, pages 301–312. Springer.

Boucheham, A., Batouche, M., and Meshoul, S. (2015b). Robust hybrid wrapper/filter biomarker discovery from gene expression data based on generalised island model. *International Journal of Computational Biology and Drug Design*, **8**(3), 251–274.

Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in bioinformatics*, **10**(5), 556–568.

Brassard, G. and Bratley, P. (1996). *Fundamentals of Algorithmics*. Prentice-Hall, Inc.

Brayet, J., Zehraoui, F., Jeanson-Leh, L., *et al.* (2014). Towards a piRNA prediction using multiple kernel fusion and support vector machine. *Bioinformatics*, **30**(17), i364–i370.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, **573**(1), 83–92.

Brent, M. R. and Guigo, R. (2004). Recent advances in gene structure prediction. *Current opinion in structural biology*, **14**(3), 264–272.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, **2**(2), 121–167.

Cadenas, J. M., Garrido, M. C., and MartíNez, R. (2013). Feature subset selection filter–wrapper based on low quality data. *Expert Systems with Applications*, **40**(16), 6241–6252.

Cantu-Paz, E. (2000). *Efficient and accurate parallel genetic algorithms*, volume 1. Springer Science & Business Media.

Chan, W.-C., Ho, M.-R., Li, S.-C., *et al.* (2012). MetaMirClust: discovery of miRNA cluster patterns using a data-mining approach. *Genomics*, **100**(3), 141–148.

Claverie, J.-M. and Notredame, C. (2011). *Bioinformatics for dummies*. John Wiley & Sons.

Cornuéjols, A. and Miclet, L. (2011). *Apprentissage artificiel: concepts et algorithmes*. Editions Eyrolles.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2009). Learning non-linear combinations of kernels. In *Advances in neural information processing systems*, pages 396–404.

Cristofanilli, M., Budd, G. T., Ellis, M. J., Stopeck, A., Matera, J., Miller, M. C., Reuben, J. M., Doyle, G. V., Allard, W. J., Terstappen, L. W., *et al.* (2004). Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *New England Journal of Medicine*, **351**(8), 781–791.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, **1**(3), 131–156.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, **7**(1), 1.

Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, **3**(02), 185–205.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* (2012). Landscape of transcription in human cells. *Nature*, **489**(7414), 101–108.

Doak, J. (1992). *An evaluation of feature selection methods and their application to computer security*. University of California, Computer Science.

Egmont-Petersen, M., Talmon, J. L., Hasman, A., and Ambergen, A. W. (1998). Assessing the importance of features for multi-layer perceptrons. *Neural networks*, **11**(4), 623–635.

Esteller, M. (2011). Non-coding RNAs in human disease. *Nature Reviews Genetics*, **12**(12), 861–874.

Fatica, A. and Bozzoni, I. (2014). Long non-coding rnas: new players in cell differentiation and development. *Nature Reviews Genetics*, **15**(1), 7–21.

Ferri, F., Pudil, P., Hatef, M., and Kittler, J. (1994). Comparative study of techniques for large-scale feature selection. *Pattern Recognition in Practice IV*, pages 403–413.

Fox, R. J. and Dimmic, M. W. (2006). A two-sample bayesian t-test for microarray data. *BMC bioinformatics*, **7**(1), 126.

García-Nieto, J. and Alba, E. (2012). Parallel multi-swarm optimizer for gene selection in dna microarrays. *Applied Intelligence*, **37**(2), 255–266.

George, G. and Raj, V. C. (2011). Review on feature selection techniques and the impact of svm for cancer classification using gene expression profile. *arXiv preprint arXiv:1109.1062*.

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., and De Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, **22**(14), e184–e190.

Ghorai, S., Mukherjee, A., Sengupta, S., and Dutta, P. K. (2011). Cancer classification from gene expression data by nppc ensemble. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **8**(3), 659–671.

Ghosh, D. and Chinnaiyan, A. M. (2005). Classification and selection of biomarkers in genomic data using lasso. *BioMed Research International*, **2005**(2), 147–154.

Goble, C. and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, **41**(5), 687–693.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, **286**(5439), 531–537.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, **12**, 2211–2268.

Gonzalez de Castro, D., Clarke, P., Al-Lazikani, B., and Workman, P. (2013). Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. *Clinical Pharmacology & Therapeutics*, **93**(3), 252–259.

Greenspan, G. and Geiger, D. (2004). High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics*, **20**(suppl 1), i137–i144.

Guan, D., Yuan, W., Lee, Y.-K., Najeebullah, K., and Rasel, M. K. (2014). A review of ensemble learning based feature selection. *IETE Technical Review*, **31**(3), 190–198.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, **3**, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, **46**(1-3), 389–422.

Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning, proceedings of 7th intentional conference on machine learning, stanford university.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato.

Handl, J., Kell, D. B., and Knowles, J. (2007). Multiobjective optimization in bioinformatics and computational biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **4**(2), 279–292.

Hassanien, A. E., Al-Shammari, E. T., and Ghali, N. I. (2013). Computational intelligence techniques in bioinformatics. *Computational biology and chemistry*, **47**, 37–47.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, **27**(2), 83–85.

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, **6**(12), e28210.

Haykin, S. (1999). Multilayer perceptrons. *Neural Networks: A Comprehensive Foundation*, **2**, 156–255.

He, Z. and Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational biology and chemistry*, **34**(4), 215–225.

Hertel, J., Hofacker, I. L., and Stadler, P. F. (2008). SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**(2), 158–164.

Hinrichs, A. S., Karolchik, D., Baertsch, R., *et al.* (2006). The UCSC genome browser database: update 2006. *Nucleic acids research*, **34**(suppl 1), D590–D598.

Hirano, T., Iwasaki, Y. W., Lin, Z., *et al.* (2014). Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. *rna*, **20**(8), 1223–1237.

Holloway, D. T., Kon, M., and DeLisi, C. (2007). Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Systems and synthetic biology*, **1**(1), 25–46.

Hu, H., Li, J., Wang, H., and Daggard, G. (2006). Combined gene selection methods for microarray data analysis. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 976–983. Springer.

Huang, C.-L. (2009). Aco-based hybrid classification system with feature subset selection and model parameters optimization. *Neurocomputing*, **73**(1), 438–448.

Huang, Y., Zhang, J. L., Yu, X. L., Xu, T. S., Wang, Z. B., and Cheng, X. C. (2013). Molecular functions of small regulatory noncoding rna. *Biochemistry (Moscow)*, **78**(3), 221–230.

Inza, I., Larrañaga, P., Blanco, R., and Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial intelligence in medicine*, **31**(2), 91–103.

Izzo, D., Ruciński, M., and Biscani, F. (2012). The generalized island model. In *Parallel Architectures and Bioinspired Algorithms*, pages 151–169. Springer.

Jaakkola, T., Diekhans, M., and Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. In *ISMB*, volume 99, pages 149–158.

Jafari, P. and Azuaje, F. (2006). An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, **6**(1), 1.

Jain, A., Vishwanathan, S., and Varma, M. (2012). SPF-GMKL: generalized multiple kernel learning with a million kernels. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 750–758. ACM.

Jain, K. K. (2005). Personalised medicine for cancer: from drug development into clinical practice.

Jain, K. K. (2010). *The handbook of biomarkers*. Springer.

Jain, K. K. (2015). *Textbook of personalized medicine, Second Edition*. Springer.

Jirapech-Umpai, T. and Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC bioinformatics*, **6**(1), 148.

Jung, I., Park, J. C., and Kim, S. (2014). piClust: a density based piRNA clustering algorithm. *Computational biology and chemistry*, **50**, 60–67.

Kahraman, C., Kaya, İ., and Çinar, D. (2010). Computational intelligence: past, today, and future. In *Computational Intelligence in Complex Decision Systems*, pages 1–46. Springer.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., *et al.* (2007). Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, **316**(5830), 1484–1488.

Keedwell, E. and Narayanan, A. (2005). *Intelligent bioinformatics: The application of artificial intelligence techniques to bioinformatics problems*. John Wiley & Sons.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., *et al.* (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, **7**(6), 673–679.

Khoshgoftaar, T. M., Fazelpour, A., Wang, H., and Wald, R. (2013). A survey of stability analysis of feature subset selection techniques. In *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on*, pages 424–431. IEEE.

Kim, J. and Kim, H. (2012). Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3. *ILAR Journal*, **53**(3-4), 232–239.

Kim, Y. S., Maruvada, P., and Milner, J. A. (2008). Metabolomics in biomarker discovery: future uses for cancer prevention.

Kittler, J. *et al.* (1978). Feature set search algorithms. *Pattern recognition and signal processing*, pages 41–60.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, **97**(1), 273–324.

Kohlmann, A., Schoch, C., Schnittger, S., Dugas, M., Hiddemann, W., Kern, W., and Haferlach, T. (2004). Pediatric acute lymphoblastic leukemia (all) gene expression signatures classify an independent cohort of adult all patients. *Leukemia*, **18**(1), 63–71.

Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review*, **35**(3), 223–240.

Kozomara, A. and Griffiths-Jones, S. (2013). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, page gkt1181.

Lakshmi, S. S. and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic acids research*, **36**(suppl 1), D173–D177.

Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, **5**, 27–72.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lanza, M., Barriuso, I., Valle, L., Domingo, M., Pérez, J., Basterrechea, J., *et al.* (2011). Comparison of different pso initialization techniques for high dimensional search space problems: A test with fss and antenna arrays. In *Antennas and Propagation (EUCAP), Proceedings of the 5th European Conference on*, pages 965–969. IEEE.

Laslett, D. and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, **32**(1), 11–16.

Le Thomas, A., Tóth, K. F., and Aravin, A. A. (2014). To be or not to be a piRNA: genomic origin and processing of piRNAs. *Genome Biol*, **15**(1), 204.

Lee, Y. and Lee, C.-K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**(9), 1132–1139.

Leung, Y. and Hung, Y. (2010). A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **7**(1), 108–117.

Li, L., Weinberg, C. R., Darden, T. A., and Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, **17**(12), 1131–1142.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**(15), 2429–2437.

Li, X. Z., Roy, C. K., Dong, X., *et al.* (2013). An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Molecular cell*, **50**(1), 67–81.

Liu, H. and Motoda, H. (2007). *Computational methods of feature selection*. CRC Press.

Liu, H. and Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, **17**(4), 491–502.

Liu, H., Liu, L., and Zhang, H. (2010). Ensemble gene selection by grouping for microarray data classification. *Journal of biomedical informatics*, **43**(1), 81–87.

Liu, J. and Zhou, H.-b. (2003). Tumor classification based on gene microarray data and hybrid learning method. In *Machine Learning and Cybernetics, 2003 International Conference on*, volume 4, pages 2275–2280. IEEE.

Liu, Q., Chen, C., Zhang, Y., and Hu, Z. (2011). Feature selection for support vector machines with rbf kernel. *Artificial Intelligence Review*, **36**(2), 99–115.

Liu, X., Krishnan, A., and Mondry, A. (2005). An entropy-based gene selection method for cancer classification using microarray data. *BMC bioinformatics*, **6**(1), 1.

Liu, X., Ding, J., and Gong, F. (2014). piRNA identification based on motif discovery. *Mol. BioSyst.*, **10**(12), 3075–3080.

Lodish, M. H., Berk, M. A., and Matsudaira, P. (2005). *Biologie moléculaire de la cellule*. De Boeck Supérieur.

López-Bigas, N. and Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic acids research*, **32**(10), 3108–3114.

Lu, Y., Tian, Q., Sanchez, M., Neary, J., Liu, F., and Wang, Y. (2007). Learning microarray gene expression data by hybrid discriminant analysis. *MultiMedia, IEEE*, **14**(4), 22–31.

Ludwig, J. A. and Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer*, **5**(11), 845–856.

Lundblad, R. L. (2010). *Development and application of biomarkers*. CRC Press.

Ma, S. and Huang, J. (2005). Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics*, **21**(24), 4356–4362.

Mäbert, K., Cojoc, M., Peitzsch, C., Kurth, I., Souchelnytskyi, S., and Dubrovska, A. (2014). Cancer biomarker discovery: current status and future perspectives. *International journal of radiation biology*, **90**(8), 659–677.

Makarova, J. and Kramerov, D. (2007). Noncoding rnas. *Biochemistry (Moscow)*, **72**(11), 1161–1178.

Mao, Z., Cai, W., and Shao, X. (2013). Selecting significant genes by randomization test for cancer classification using gene expression data. *Journal of biomedical informatics*, **46**(4), 594–601.

Martinez, E., Alvarez, M. M., and Trevino, V. (2010). Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. *Computational biology and chemistry*, **34**(4), 244–250.

Matsui, S. (2013). Genomic biomarkers for personalized medicine: development and validation in clinical studies. *Computational and mathematical methods in medicine*, **2013**.

Matthiesen, R. (2010). *Bioinformatics methods in clinical research*. Springer.

Menor, M. S., Baek, K., and Poisson, G. (2015). Prediction of Mature MicroRNA and Piwi-Interacting RNA without a Genome Reference or Precursors. *International journal of molecular sciences*, **16**(1), 1466–1481.

Mitra, S. and Acharya, T. (2005). *Data mining: multimedia, soft computing, and bioinformatics*. John Wiley & Sons.

Mitra, S. and Hayashi, Y. (2006). Bioinformatics with soft computing. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, **36**(5), 616–635.

Murray, D., Doran, P., MacMathuna, P., and Moss, A. C. (2007). In silico gene expression analysis–an overview. *Molecular Cancer*, **6**(1), 50.

Nagaraj, N. S. (2009). Evolving omics technologies for diagnostics of head and neck cancer. *Briefings in Functional Genomics*, page elp004.

Nair, M., Singh Sandhu, S., and K Sharma, A. (2014). Prognostic and predictive biomarkers in cancer. *Current cancer drug targets*, **14**(5), 477–504.

Namsrai, E., Munkhdalai, T., Li, M., Shin, J.-H., Namsrai, O.-E., and Ryu, K. H. (2013). A feature selection-based ensemble method for arrhythmia classification. *Journal of Information Processing Systems*, **9**(1), 31–40.

Narendra, P. M. and Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *Computers, IEEE Transactions on*, **100**(9), 917–922.

Network, R. A. (2010). Biomarkers in cancer: an introductory guide for advocates. récupéré le 26 novembre 2015.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K.-W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational biology*, **8**(1), 37–52.

Ng, M. and Chan, L. (2005). Informative gene discovery for cancer classification from microarray expression data. In *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, pages 393–398. IEEE.

O'Brien, S. G., Guilhot, F., Larson, R. A., Gathmann, I., Baccarani, M., Cervantes, F., Cornelissen, J. J., Fischer, T., Hochhaus, A., Hughes, T., *et al.* (2003). Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *New England Journal of Medicine*, **348**(11), 994–1004.

Ohler, U., Yekta, S., Lim, L. P., *et al.* (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *Rna*, **10**(9), 1309–1322.

Okun, O. and Skarlas, L. (2011). *Feature selection and ensemble methods for bioinformatics: algorithmic classification and implementations*. Medical Information Science Reference.

Osl, M., Netzer, M., Dreiseitl, S., and Baumgartner, C. (2012). *Applied data mining: From biomarker discovery to decision support systems*. Springer.

Quackenbush, J. (2006). Microarray analysis and tumor classification. *New England Journal of Medicine*, **354**(23), 2463–2472.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, **1**(1), 81–106.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Rätsch, G., Sonnenburg, S., Srinivasan, J., Witte, H., Müller, K.-R., Sommer, R.-J., and Schölkopf, B. (2007). Improving the caenorhabditis elegans genome annotation using machine learning. *PLoS Comput Biol*, **3**(2), e20.

Rice, P., Longden, I., Bleasby, A., *et al.* (2000). Emboss: the european molecular biology open software suite. *Trends in genetics*, **16**(6), 276–277.

Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W., and Moore, J. H. (2003). Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC bioinformatics*, **4**(1), 28.

Rittner, D. and McCabe, T. L. (2009). *Encyclopedia of biology*. Infobase Publishing.

Rose, N. R. and Klose, R. J. (2014). Understanding the relationship between dna methylation and histone lysine methylation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1839**(12), 1362–1372.

Saeys, Y., Degroeve, S., Aeyels, D., Rouzé, P., and Van de Peer, Y. (2004). Feature selection for splice site prediction: a new method using eda-based feature ranking. *BMC bioinformatics*, **5**(1), 64.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, **23**(19), 2507–2517.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Machine learning and knowledge discovery in databases*, pages 313–325. Springer.

Soldà, G., Makunin, I. V., Sezerman, O. U., *et al.* (2009). An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. *Briefings in bioinformatics*, **10**(5), 475–489.

Somorjai, R. L., Dolenko, B., and Baumgartner, R. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **19**(12), 1484–1491.

Srinivas, P. R., Verma, M., Zhao, Y., and Srivastava, S. (2002). Proteomics for cancer biomarker discovery. *Clinical chemistry*, **48**(8), 1160–1169.

Sumathi, S. and Paneerselvam, S. (2010). *Computational intelligence paradigms: theory & applications using MATLAB*. CRC Press.

Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., and Wang, K. (2013). Feature selection using dynamic weights for classification. *Knowledge-Based Systems*, **37**, 541–549.

Sweilam, N. H., Tharwat, A., and Moniem, N. A. (2010). Support vector machine for diagnosis cancer disease: A comparative study. *Egyptian Informatics Journal*, **11**(2), 81–92.

Tafer, H. and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics*, **24**(22), 2657–2663.

Tang, Y., Zhang, Y.-Q., and Huang, Z. (2007). Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **4**(3), 365–381.

Tatusova, T. A. and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS microbiology letters*, **174**(2), 247–250.

Tempel, S. and Tahi, F. (2012). A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic acids research*, **40**(11), e80–e80.

Terpos, E., Dimopoulos, M. A., Shrivastava, V., Leitzel, K., Christoulas, D., Migkou, M., Gavriatopoulou, M., Anargyrou, K., Hamer, P., Kastritis, E., *et al.* (2010). High levels of serum timp-1 correlate with advanced disease and predict for poor survival in patients with multiple myeloma treated with novel agents. *Leukemia research*, **34**(3), 399–402.

Thomas, J. G., Olson, J. M., Tapscott, S. J., and Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, **11**(7), 1227–1236.

Thomson, T. and Lin, H. (2009). The biogenesis and function PIWI proteins and pirnas: progress and prospect. *Annual review of cell and developmental biology*, **25**, 355.

Tran, V. D. T., Tempel, S., Zerath, B., *et al.* (2015). miRBoost: boosting support vector machines for microRNA precursor classification. *RNA*, **21**(5), 775–785.

Upadhyaya, S. R. (2013). Parallel approaches to machine learning- a comprehensive survey. *Journal of Parallel and Distributed Computing*, **73**(3), 284–292.

Valentini, G., Tagliaferri, R., and Masulli, F. (2009). Computational intelligence and machine learning in bioinformatics. *Artificial intelligence in medicine*, **45**(2), 91–96.

Van Hulse, J., Khoshgoftaar, T. M., Napolitano, A., and Wald, R. (2012). Threshold-based feature selection techniques for high-dimensional bioinformatics data. *Network modeling analysis in health informatics and bioinformatics*, **1**(1-2), 47–61.

Vapnik, V. (1998). Statistical learning theory. 1998.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Varma, M. and Babu, B. R. (2009). More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome. *science*, **291**(5507), 1304–1351.

Vourekas, A., Zheng, K., Fu, Q., *et al.* (2015). The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. *Genes & development*, **29**(6), 617–629.

Wang, K., Liang, C., Liu, J., *et al.* (2014a). Prediction of piRNAs using transposon interaction and a support vector machine. *BMC bioinformatics*, **15**(1), 419.

Wang, L., Wang, X., Arkin, A. P., and Samoilov, M. S. (2013). Inference of gene regulatory networks from genome-wide knockout fitness data. *Bioinformatics*, **29**(3), 338–346.

Wang, W., Yoshikawa, M., Han, B. W., *et al.* (2014b). The initial uridine of primary piRNAs does not create the tenth adenine that Is the hallmark of secondary piRNAs. *Molecular cell*, **56**(5), 708–716.

Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., and Mewes, H. W. (2005). Gene selection from microarray data for cancer classification a machine learning approach. *Computational biology and chemistry*, **29**(1), 37–46.

Washietl, S., Will, S., Hendrix, D. A., *et al.* (2012). Computational analysis of noncoding RNAs. *Wiley Interdisciplinary Reviews: RNA*, **3**(6), 759–778.

Weick, E.-M. and Miska, E. A. (2014). piRNAs: from biogenesis to function. *Development*, **141**(18), 3458–3471.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, K.-P. and Wang, S.-D. (2009). Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. *Pattern Recognition*, **42**(5), 710–717.

Wu, M.-Y., Dai, D.-Q., Shi, Y., Yan, H., and Zhang, X.-F. (2012). Biomarker identification and cancer classification based on microarray data using laplace naive bayes model with mean shrinkage. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **9**(6), 1649–1662.

Xing, E. P., Jordan, M. I., Karp, R. M., *et al.* (2001). Feature selection for high-dimensional genomic microarray data. In *ICML*, volume 1, pages 601–608. Citeseer.

Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press.

Xu, J., Sun, L., Gao, Y., and Xu, T. (2013). An ensemble feature selection technique for cancer recognition. *Bio-medical materials and engineering*, **24**(1), 1001–1008.

Xue, B., Zhang, M., and Browne, W. N. (2012). Multi-objective particle swarm optimisation (pso) for feature selection. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 81–88. ACM.

Yamanaka, S., Siomi, M. C., and Siomi, H. (2014). piRNA clusters and open chromatin structure. *Mobile DNA*, **5**(1), 22.

Yang, P., Hwa Yang, Y., B Zhou, B., and Y Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, **5**(4), 296–308.

Yang, P., Liu, W., Zhou, B. B., Chawla, S., and Zomaya, A. Y. (2013). Ensemble-based wrapper methods for feature selection and class imbalance learning. In *Advances in Knowledge Discovery and Data Mining*, pages 544–555. Springer.

Yazdani, S., Shanbehzadeh, J., and Aminian, E. (2013). Feature subset selection using constrained binary/integer biogeography-based optimization. *ISA transactions*, **52**(3), 383–390.

Yeung, K. Y., Bumgarner, R. E., *et al.* (2003). Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome biology*, **4**(12), R83–R83.

Yu, B., Cassani, M., Wang, M., *et al.* (2015). Structural insights into Rhino-mediated germline piRNA cluster formation. *Cell research*.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, **5**, 1205–1224.

Yu, L., Han, Y., and Berens, M. E. (2012). Stable gene selection from microarray data via sample weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **9**(1), 262–272.

Zhang, H., Wang, H., Dai, Z., Chen, M.-s., and Yuan, Z. (2012). Improving accuracy for cancer classification with a new algorithm for genes selection. *BMC bioinformatics*, **13**(1), 1.

Zhang, P., Si, X., Skogerbø, *et al.* (2014). piRBase: a web resource assisting piRNA functional study. *Database: the journal of biological databases and curation*, **2014**.

Zhang, X., Shi, L., Chen, G., and Yap, Y. L. (2011a). Integrative omics technologies in cancer biomarker discovery. *Omics Technologies in Cancer Biomarker Discovery*, **129**.

Zhang, Y., Wang, X., and Kang, L. (2011b). A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics*, **27**(6), 771–776.

Zhu, S., Wang, D., Yu, K., Li, T., and Gong, Y. (2010a). Feature selection for gene expression using model-based entropy. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **7**(1), 25–36.

Zhu, Z., Ong, Y.-S., and Zurada, J. M. (2010b). Identification of full and partial class relevant genes. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **7**(2), 263–277.

Zurada, J., Marks, R., and Robinson, J. (1995). Review of computational intelligence: imitating life.