UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MÉTHODES ET PLATEFORMES BIOINFORMATIQUES BASÉES SUR L'APPRENTISSAGE AUTOMATIQUE POUR LA CLASSIFICATION EFFICACE DE SÉQUENCES BIOLOGIQUES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

MOHAMED AMINE REMITA

AVRIL 2017

REMERCIEMENTS

Je remercie mon directeur de recherche Abdoulaye Baniré Diallo. Je suis reconnaissant à Abdoulaye pour son soutien, son implication, son aide et ses conseils précieux. Je trouve en lui un professeur, un guide, un inspirateur, un ami et un proche. C'est avec un grand plaisir que je compte poursuivre mes études au doctorat sous sa supervision.

Je remercie mes ami-e-s proches et membres du laboratoire de bioinformatique Golrokh Kiani, Bruno Daigle et Ahmed Halioui.

Je remercie Fathey Sarhan et son associée de recherche Zahra Agharbaoui, du département des sciences biologiques de l'UQAM, pour m'avoir impliqué dans leur projet de recherche et soutenu dans mes demandes de bourses.

Je remercie mes collaboratrices et collaborateurs de mes projets de recherche Abdoulaye Baniré Diallo, Fathey Sarhan, Zahra Agharbaoui, Mickael Leclercq, Etienne Lord, Mohamed Badawi, Mario Houde, Jean Danyluk, Ahmed Halioui, Abou Abdallah Malick Diouara, Bruno Daigle et Golrokh Kiani.

Je remercie mes collègues du laboratoire de bioinformatique ainsi les anciens membres du laboratoire Alix boc, Etienne Lord, Mickael Leclercq et Dunarel Badesco.

Je remercie mes professeur-e-s de la maîtrise en informatique Louise Laforest, Étienne Gagnon, Mohamed Bouguessa, Éric Beaudry et Petko Valtchev.

Je remercie Vladimir Makarenkov et Anne Bergeron professeur-e-s de la bioinformatique au département d'informatique.

Je remercie le corps administratif du département d'informatique notamment Marie-Claude Coté, Sylvie Dorval et Kathleen Jackson.

Je remercie le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), le Fonds de recherche du Québec – Nature et technologies (FRQNT) et l'Université du Québec à Montréal (UQAM) pour le soutien financier (bourse d'études supérieures du Canada Alexander Graham Bell, bourse de maîtrise du FRQNT et bourse d'excellence de la Faculté des sciences, respectivement).

Je remercie Abdel Aziz Hadj Moussa, Rastin Azizbigloo et Khaoula Mazouzi et tous mes ami-e-s pour les bons moments qu'on a passés ensemble.

Je remercie ma chère mère Bariza et mon cher père Ferhat ainsi mes chères sœurs Imene et Wissem Elkhouloud.

DÉDICACES

À ma grand-mére Tata Saâda

À ma mère Bariza

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ALGORITHMES

RÉSUMÉ

Les récentes avancées des technologies de séquençage des biomolécules ont contribué à la génération de quantités colossales de séquences et de données biologiques. La classification de ces séquences est importante dans les différentes analyses de séquences telles que l'identification des éléments génomiques, la prédiction de leurs caractéristiques et fonctions, et l'inférence de leurs relations phylogénétiques et taxonomiques. Les méthodes de classification assignent une nouvelle séquence à un ensemble de séquences de référence qui partagent des propriétés et des traits similaires. Ces méthodes appartiennent à plusieurs catégories d'approches qui sont basées sur l'alignement de séquences ou la phylogénie, ou indépendantes de l'alignement de séquences. Plusieurs méthodes de classification sont généralement spécifiques à un type d'organisme et ne sont pas adaptées à un volume énorme de séquences. Dans ce mémoire, nous abordons deux problématiques de classification des séquences nucléotidiques, la classification des séquences virales en rangs taxonomiques et l'identification des microARNs à partir d'un lot de séquences de petits ARNs. Nous avons développé des méthodes de classification intégratives, indépendantes de l'alignement de séquences et basées sur l'apprentissage automatique. Ces nouvelles méthodes sont génériques, adaptées à plusieurs types d'organismes, et peuvent traiter des grandes quantités de données de nature hétérogène. Les méthodes développées sont efficaces, rapides et ont des performances comparables à celles des méthodes existantes. Finalement, elles sont implémentées dans des plateformes web publiques facilitant la réutilisation et la reproductibilité des expériences de classification par la communauté.

Mots clés : Classification des séquences nucléotidiques, apprentissage automatique, prédiction, classification des séquences virales, identification des microARNs, plateformes intégratives.

INTRODUCTION

Dans le domaine de la *bioinformatique*, on conçoit des algorithmes et des programmes informatiques et statistiques pour la gestion, le traitement, l'analyse et la modélisation des données afin de résoudre des problèmes en sciences de la vie tels que la biologie, l'agronomie et la médecine. La *bioinformatique* est multi-disciplinaire qui implique la biologie, l'informatique, les statistiques et les mathématiques. Les données biologiques sont constituées, de façon générale mais non exhaustive, des séquences, des structures, des fonctions et des interactions des biomolécules.

L'analyse de séquences biologiques est l'une des pratiques pionnières de la bioinformatique. Elle permet l'inférence d'homologie, l'annotation fonctionnelle et structurelle ainsi l'identification des variations et marqueurs génétiques. Les approches d'analyse de séquences biologiques ont vu le jour avec la génération des premiers fragments de séquences des protéines et des acides nucléiques dans les années 1960 et 1970. Initialement, ces approches étaient basées sur l'algorithmique du texte (domaine de l'informatique qui traite les chaines de caractères). Cela inclut l'alignement, la comparaison et le regroupement des séquences, le calcul de distance entre deux séquences, la recherche des motifs et l'assemblage des sous-séquences, etc. Durant les deux décennies suivantes (1980 - 1990), les séquences biologiques de différentes espèces sont devenues plus abondantes grâce à la maîtrise et à l'automatisation des techniques de séquençage de première génération. Des nouvelles méthodes ont été développées pour analyser ces séquences et de nouveaux concepts ont été introduits tels que la modélisation probabiliste (chaînes de Markov) des séquences et des familles de gènes. Des méthodes issues

de l'apprentissage automatique (*machine learning*), telles que les réseaux de neurones artificiels (*ANN*), les séparateurs à vaste marge (*SVM*) et les modèles de Markov cachées (*HMM*), ont été utilisées dans le but d'identifier et classer les séquences. Également, des approches avancées ont été développées pour les modèles d'évolution de l'ADN ainsi pour la construction phylogénétique (telles que le maximum de vraisemblance et l'inférence bayésienne). Dans les années 2000, le développement de nouvelles technologies de biologie moléculaire, particulièrement le séquençage à haut débit, a facilité le séquençage des transcriptomes et des génomes des milliers d'espèces (animales, végétales et microbiennes). Ces technologies génèrent des quantités massives de données biologiques (de l'ordre de téraoctets). L'adaptation des méthodes d'analyses existantes et le développement des techniques modernes sont devenus nécessaires afin de manipuler, traiter et interpréter un volume important de données génomiques de nature complexe et hétérogène. Plusieurs méthodes d'apprentissage automatique (supervisées et non supervisées) sont appliquées de plus en plus dans l'analyse de séquences biologiques en raison de leur rapidité, la possibilité de leur interprétation, leur adaptation et extension, et la disponibilité de leurs implémentations informatiques.

Le présent mémoire s'intéresse à la classification de séquences biologiques, l'un des domaines importants de l'analyse de séquences. La classification de séquences assigne une nouvelle séquence à un ensemble de séquences connues partageant des propriétés, des caractéristiques, des traits ou des fonctions similaires. Elle sert à prédire des gènes codants et non codants, à annoter des éléments et des régions génomiques ainsi à identifier des nouvelles souches et espèces. Les expériences de la classification traitent des données génétiques, génomiques et métagénomiques.

Les approches de classification peuvent être dédiées (spécifiques) à un élément génétique (promoteurs des gènes, sites d'épissage, une famille de gènes, petits ARNs, etc.), un type de tissu (hépatique, nerveux, etc.), une maladie, un dysfonctionne-

ment (cancers, diabète, etc.), une espèce ou un rang taxonomique (Homo sapiens, Arabidopsis, plantes, retrovirus, etc.). Cependant, il existe des approches génériques (universelles) conçues pour la classification des séquences dans plusieurs types de problèmes (classification des gènes codants, des viromes ou des données métagénomiques, etc.). Dans les approches algorithmiques de la classification de séquences, il existe deux catégories de méthodes. 1) La première catégorie regroupe des méthodes basées sur un alignement de séquences, comme la recherche par similarité, les comparaisons par paire et les inférences phylogénétiques. 2) La deuxième catégorie est indépendante de l'alignement de séquences (*alignment-free*). Ces méthodes transforment les séquences ou leurs relations en des vecteurs. Les vecteurs sont ensuite utilisés pour entrainer un modèle statistique ou d'apprentissage automatique. Sous une autre perspective, les approches de classification peuvent être individuelles utilisant une seule méthode de classification, comme elles peuvent être intégratives combinant plusieurs méthodes et algorithmes de classification.

Les travaux de recherche décrits dans ce mémoire portent sur ces différents aspects de la classification de séquences en mettant l'accent sur les méthodes basées sur l'apprentissage automatique (les autres méthodes sont aussi utilisées ou évaluées). Le mémoire inclut un chapitre d'introduction au contexte et des notions abordées (en biologie et en apprentissage automatique), trois principaux chapitres ainsi qu'une conclusion. Les trois principaux chapitres sont présentés sous forme d'articles scientifiques acceptés ou publiés dans des journaux avec comités de lecture.

— Article 1 (chapitre 2) :

  Remita M.A., Halioui A., Diouara A.A.M., Daigle B., Kiani G. and Diallo A.B. *A machine learning approach for viral genome classification.* Accepté le 15 mars 2017 à **BMC Bioinformatics**.

  L'article présente une nouvelle méthode ainsi une plateforme générique

pour la classification des séquences virales. Cette méthode est indépendante de l'alignement, inspirée d'une technique de biologie moléculaire et est basée sur des algorithmes d'apprentissage supervisé.
J'ai une contribution majeure dans ce projet. La conception de l'approche est faite avec Ahmed Halioui sous la supervision de Abdoulaye Baniré Diallo. J'ai développé et implémenté la méthode (programme principal), la plateforme web (back-end et front-end) et la base de données. J'ai construit les jeux de données et réalisé les simulations et les expériences de classification. Les coauteurs ont participé dans les différentes étapes du projet.

— Article 2 (chapitre 3) :
Agharbaoui Z., Leclercq M., Remita M.A., Badawi M.A, Lord E., Houde M., Danyluk J., Diallo A.B. and Sarhan F. (2015). *An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress.* **BMC Genomics**, 16(1), 339.
L'article décrit une approche intégrative pour la classification des microARNs appliquée au génome du blé. Cette approche, conçue pour la prédiction des microARNs des plantes, combine plusieurs étapes de traitement de données et de multiples algorithmes de prédiction.
J'ai une contribution principale, avec Mickael Leclercq, dans l'aspect bioinformatique du projet sous la supervision de Abdoulaye Baniré Diallo. J'ai participé à la conception et le développement de l'approche bioinformatique, l'intégration des données, la génération et l'agrégation des résultats, leur analyse et leur interprétation.

— Article 3 (chapitre 4) :
Remita M.A., Lord E., Agharbaoui Z., Leclercq M., Badawi M.A, Sarhan F. and Diallo A.B. (2016) *A novel comprehensive wheat miRNA database,*

including related bioinformatics software. **Current Plant Biology**, 7, 31-33.

L'article présente une plateforme web implémentant des méthodes utilisées dans la prédiction des microARNs. La plateforme offre aussi une ressource de données de microARNs associées à différentes conditions de développement du blé.

J'ai conçu et réalisé, conjointement avec Etienne Lord, la plateforme et la base de données WMP sous la direction de Abdoulaye Baniré Diallo.

CHAPITRE I

NOTIONS DE BIOLOGIE ET D'APPRENTISSAGE AUTOMATIQUE

1.1 Notions de biologie

1.1.1 Acide nucléique

Un acide nucléique est une macromolécule (polymère) constituée d'un empilement de plusieurs monomères appelés nucléotides. Chaque nucléotide est composé de trois éléments : un sucre à 5 carbones (pentose), un groupe phosphate et une base nucléique. Les bases nucléiques sont des molécules organiques azotées au nombre de cinq principalement, qui sont l'adénine (A), la cytosine (C), la guanine (G), la thymine (T) et l'uracile (U). Il existe deux types d'acides nucléiques : l'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN). L'ADN héberge le génome, l'ensemble de l'information génétique d'un organisme. Il contient des gènes codants (transcrits en ARN messagers), des gènes non codants (transcrits en ARN non codants ou éléments transposables) et des régions non transcrites. Il est formé de deux brins antiparallèles (orientés en sens opposé) et enroulés l'un autour de l'autre (en double hélice). Les ARNs sont divisés en deux catégories : ARNs codants et ARNs non codants. L'ARN codant est appelé ARN messager (ARNm), une copie simple brin de l'ADN qui sera traduite en protéine. Les ARNs non codants s'impliquent dans différents mécanismes cellulaires tels que la traduction protéique (ARN ribosomique (ARNr) et ARN de transfert (ARNt)), la catalyse

des réaction chimique (Ribozyme) et la régulation de l'expression génique (ARN interférent (ARNi)).

1.1.2 microARN

Un des types des ARNi est le microARN (miARN). C'est un petit simple brin d'ARN d'une taille de 18 à 25 nucléotides. Les miARNs sont des régulateurs post-transcriptionnels de l'expression des gènes et sont impliqués dans plusieurs fonctions physiologiques et voies métaboliques (développement, croissance, stress biotique et abiotique, cancer et apoptose). Chaque miARN est issu d'une séquence, d'une centaine de nucléotides appelée pré-miARN, qui se replie en épingle à cheveux (structure tige-boucle) comprenant une complémentarité imparfaite (boucle et hernies) et formant un duplexe miARN 5p - miARN 3p (figure 1.1). Les gènes codant les miARNs se localisent dans plusieurs régions du génome et peuvent former des clusters de miARNs. On les trouve chez les animaux, les plantes et les virus où leurs biogenèses et mécanismes de ciblage des gènes sont différents. Les miARNs peuvent être conservés entre plusieurs espèces du même règne comme ils peuvent être spécifique pour un clade ou une espèce (miARN jeune).

1.1.3 Peptides et protéines

Les peptides et les protéines sont des polymères linéaires composés d'acides aminés. La différence entre les deux réside dans le nombre d'acides aminés. Les protéines sont plus grandes que les peptides. Ces macromolécules sont généralement issues de la traduction d'ARNm effectuée par les ribosomes. Ces derniers assemblent les résidus d'acides aminés selon un ordre défini par une succession de codons (triplets de nucléotides) portés sur l'ARNm et traduits en acides aminés par des ARNts selon le code génétique. Vingt-deux acides aminés, dits protéinogènes, participent dans la biosynthèse des protéines. L'enchaînement des acides

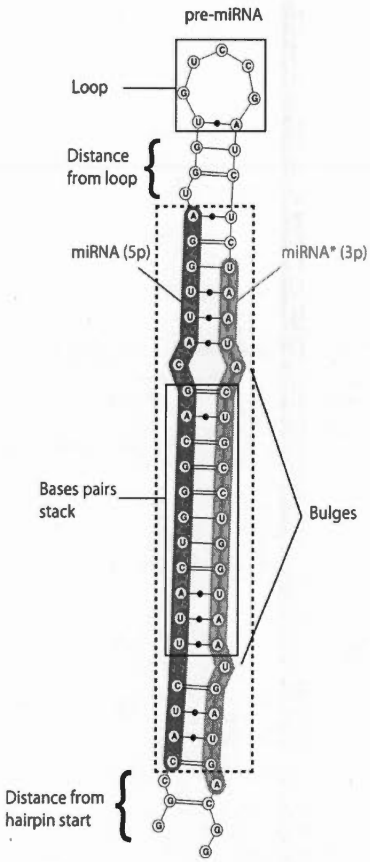


Figure 1.1: Structure secondaire d'un précurseur de miARN. Source : Leclercq *et al.* (2013)

aminés dans la protéine constitue sa structure primaire.

1.1.4 Enzyme de restriction

Une enzyme de restriction est une protéine endonucléase qui coupe l'ADN dans une région spécifique appelée site de restriction (4 à 8 paires de bases). Il existe des centaines d'enzymes de restrictions classifiées principalement en trois groupes (d'autres groupes existent) selon leurs compositions, leurs besoins en cofacteurs enzymatiques, les sites de restrictions qu'elles reconnaissent et la position du cli-

vage. Les enzymes de type I et III coupent un fragment d'ADN en un endroit distant du site de restriction, cependant les enzymes de type II coupent le fragment en endroit spécifique du site. Ces enzymes sont exprimées par des bactéries et sont impliquées dans un système de défense contre les virus (bactériophages). Ce système est appelé système de restriction/modification associant deux enzymes reconnaissant le même site de restriction, une enzyme qui clive le fragment du virus et l'autre qui méthyle le fragment de la bactérie empêchant sa coupure. Grâce à leurs propriétés, les enzymes de restriction sont utilisées dans les laboratoires de biologie moléculaire et génie génétique. Plusieurs techniques de biologie moléculaire se basent sur ces enzymes tels que le clonage moléculaire et le polymorphisme de longueur des fragments de restriction (*restriction fragment length polymorphism*, RFLP). La technique RFLP est utilisée dans la discrimination des séquences d'ADN homologues.

1.1.5 Virus

Les virus sont des agents biologiques infectant les cellules d'autres organismes pour se répliquer. Les virus dans leur forme extra-cellulaire sont des particules (appelées virions) constituées d'un génome (acide nucléique) entouré par une capside de protéines. Les virus sont des entités ubiquitaires et abondantes. Il est estimé qu'il existe 10^{31} virus sur la terre, la majorité des virus sont des bactériophages (virus infectant les bactéries) (Breitbart et Rohwer, 2005). Ils montrent aussi une grande diversité (milliers de génotypes). Leurs génomes peuvent être de l'ADN ou l'ARN ; simple ou double brin ; linéaire, circulaire, partiellement circulaire ou segmenté ; de polarité positive, négative ou ambisens.

1.1.6 Séquençage

Le séquençage est le processus de détermination de la structure primaire d'une biomolécule c'est-à-dire l'ordre linéaire de ses composants (les nucléotides pour les acides nucléiques et les acides aminés pour les protéines). Le résultats du séquençage est une séquence de symboles appartenant un alphabet spécifique à chaque type de biomolécule. Les séquences issues du séquençage peuvent être manipulées et analysées par des méthodes informatiques et statistiques.

1.2 Données utilisées en apprentissage automatique

Dans un problème d'apprentissage automatique, les données sont représentées sous la forme d'**objets** et d'**attributs**. Un jeu de données est constitué par un ensemble d'objets. Les objets sont nommés aussi *exemples, instances, enregistrements, observations, vecteurs, entités*, etc. De l'autre coté, les caractéristiques et les propriétés des objets sont décrites par un ensemble d'attributs. Les attributs sont connus aussi par les noms *dimensions, champs, features* et *variables*. Pour analyser des données, plusieurs étapes sont nécessaires dont la description statistique, le prétraitement, la classification et l'interprétation. Les sous- sections qui suivent décrivent les deux premières étapes.

1.2.1 Types d'attributs

Les attributs se distinguent en deux catégories : **qualitatifs** (catégoriels) et **quantitatifs**. Les attributs qualitatifs ont des valeurs réparties en différentes catégories dont les caractéristiques sont discrets. Par ailleurs, les attributs quantitatifs sont des nombres représentant des comptages et des mesures. Les attributs quantitatifs **discrets** ont des valeurs finies ou dénombrables contrairement aux attributs **continus**. Une autre manière de distinguer les types d'attributs est l'utilisation des propriétés des nombres. Ces propriétés sont : la distinction ($=$ et $\neq$), l'ordre

($<, \leq, >$ et $\geq$), l'addition ($+$ et $-$) et la multiplication ($*$ et $/$). Le tableau 1.1 décrit les types d'attributs selon ces propriétés ainsi les opérations statistiques valides pour ces attributs.

1.2.2 Description statistique des données

Des mesures statistiques simples peuvent être appliquées pour avoir une vue d'ensemble des jeux de données et décrire leurs différentes caractéristiques. Supposons qu'on a N objets $x_1, x_2, ..., x_i, ..., x_N$ décrits par un attribut X. On peut calculer pour chaque valeur x_i une **fréquence** qui est le nombre d'objets qui ont la valeur x_i dans l'attribut X divisé par le nombre d'objets N. Aussi, des mesures de tendance centrale et de dispersion peuvent être calculées sur l'ensemble des données.

1.2.2.1 Mesures de tendance centrale

Une mesure de tendance centrale (ou la position) est une valeur qui pointe sur le centre ou le milieu d'un jeu de données. Les mesures principales de position sont :

- **Le mode** est la valeur x_i la plus fréquente. Il peut être calculé pour des attributs catégoriels et numériques. Une variable peut avoir un seul mode (variable unimodale) comme plusieurs (variable plurimodale).
- **La moyenne arithmétique** ($\overline{x}$) est la mesure la plus utilisée pour décrire la position d'un ensemble de valeurs. Elle peut être simple, calculée comme suit :

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{1.1}$$

  Elle peut être une **moyenne pondérée** en associant chaque valeur x_i à un poids ou un coefficient w_i :

$$\overline{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \tag{1.2}$$

Tableau 1.1: Les différents types d'attributs. Adapté de Tan *et al.* (2006)

Type d'attribut		Description	Propriétés	Operations statistiques
Qualitatif (catégoriel)	Nominal	Les valeurs sont des noms ou des symbols différents. Avec ces attributs, on peut seulement distinguer un objet d'un autre.	Distinction	Mode, entropie, coefficient de contingence, test du χ^2
	Ordinal	Les valeurs sont des catégories ordonnées. Leurs différences n'ont pas de sens	Distinction et ordre	Médiane, quantiles, tests non paramétriques
Quantitatif (numérique)	Intervalle	Les différences entre les valeurs ont du sens mais il n'y a pas de zéro naturel de référence.	Distinction, ordre et addition	Moyenne, écart type, correlation linéaire, test de Student et test de Fisher
	Rapport	Les rapports entre les valeurs ont du sens et il y a un zéro naturel de référence	Distinction, ordre, addition et multiplication	Moyennes géométrique et moyenne harmonique

Une moyenne arithmétique peut être biaisée par des valeurs extrêmes. Ce biais peut être retiré en calculant une moyenne sans les valeurs extrèmes (**moyenne tronquée**).

— **La médiane** d'un ensemble de valeurs triées est la valeur du milieu. la médiane sépare l'ensemble en deux moitiés égales, une moitié avec des valeurs inférieures à la médiane et l'autre avec des valeurs supérieures. Cette mesure est plus descriptive de la tendance centrale que la moyenne dans le cas d'une distribution asymétrique des valeurs ou la présence de valeurs extrêmes. Elle est calculée pour les variables numériques mais peut être adaptée pour des variables ordinales.

Le mode, la moyenne et la médiane ont la même valeur dans une distribution de données symétrique (figure 1.2a). Généralement les données ne sont pas symétriques. Elles présentent soit une distribution asymétrique positive où le mode est inférieur à la médiane et la moyenne (figure 1.2b) ou négative où le mode est plus grand que les deux autres mesures (figure 1.2c).

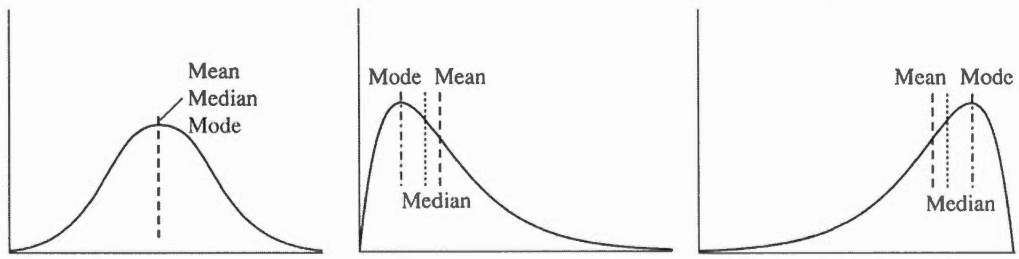


Figure 1.2: Symétrie contre asymétire positive et négative. Source : Han *et al.* (2011)

1.2.2.2 Mesures de dispersion

Les mesures de dispersion sont un moyen important pour la description des données. Elles représentent la variation et l'étendue des valeurs d'un attribut ou d'une variable. Elles indiquent si les valeurs sont largement étalées ou relativement concentrées autour d'une autre valeur telle que la moyenne. Finalement, ces mesures sont utiles pour identifier les valeurs aberrantes (outliers). Les mesures les plus utilisées sont :

- **L'étendue** d'un ensemble de valeurs est la différence entre la valeur maximale et la valeur minimale.
- **Les quantiles** sont des valeurs d'un ensemble de données ordonné qui le partitionnent en sous-ensembles consécutifs et de tailles égales. Le k^{eme} $q-quantile$ est la valeur x tel que k/q des données ont une valeur inférieure à x et $(q-k)/q$ des données ont une valeur supérieure à x, où $0 < k < q$. Les 100-quantiles, connus par **centiles**, partitionnent les données en 100 sous-ensembles. Les 4-quantiles sont les trois valeurs (Q_1, Q_2 et Q_3) qui partitionnent les données en quatre parts égales. Ils sont connus par le nom **quartiles**. La médiane est un cas particulier des quantiles (2-quantile) et elle correspond à Q_2 et à $x_{50\%}$ pour les quartiles et les centiles respectivement.
- **l'écart interquartile** est l'étendue entre le premier quartile et le troisième quartile ($Q_3 - Q1$). Cette mesure est, généralement, plus robuste aux valeurs extrêmes que l'étendue décrite auparavant.
- **La variance** et **l'écart type** sont des mesures importantes de la dispersion et la variation des valeurs autour de la moyenne. L'écart type (σ), qui est la racine carrée de la variance, représente la déviation moyenne des valeurs par rapport à la moyenne. Il a une faible valeur si l'ensemble des données est

concentré sur la moyenne, et une valeur élevée si les données sont réparties sur un grand nombre de valeurs. La formule de la variance est donnée par l'équation 1.2.2.2.

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{1.3}$$

L'écart type est toujours positif et il est nul si toutes les données ont la même valeur. Il a la même unité que les données originales. La variance et l'écart type sont sensibles aux valeurs aberrantes parce qu'ils sont calculés en utilisant la moyenne.

1.2.3 Prétraitement des données

Les grandes masses de données contiennent généralement du bruit, des inconsistances, des valeurs manquantes et des duplications. Cela peut être dû aux erreurs et limitations de mesures et aux origines multiples et hétérogènes des données. Les données de mauvaise qualité ne sont pas propices pour la réalisation d'analyses et modélisations pertinentes. L'étape de prétraitement des données consiste en plusieurs techniques qui visent à réduire ou éliminer ces imperfections. Ces techniques sont : l'intégration, le nettoyage, la réduction de la dimension, la transformation et la discrétisation (Han *et al.*, 2011; Tan *et al.*, 2006; Witten *et al.*, 2011).

1.2.3.1 Intégration

Les technique d'apprentissage automatique peuvent utiliser des jeux de données issus de plusieurs sources. La combinaison de plusieurs jeux de données peut introduire des inconsistances (par exemple le conflit de valeurs pour un attribut avec différents unités de mesures), des redondances (un attribut peut exister sous différents noms ou peut être déduit à partir d'un autre) et des duplications d'enregistrements. L'étape d'intégration doit fournir des solutions tels que la résolution

de l'hétérogénéité sémantique, analyse de corrélation et la détection des duplications.

1.2.3.2 Nettoyage

Le nettoyage des données comporte le traitement des données incomplètes. Plusieurs méthodes peuvent être utilisées pour traiter les valeurs manquantes. La plus simple est l'élimination des enregistrements contenant ces vides ou l'utilisation d'une constante globale pour remplir les vides. Une autre méthode est de remplacer les valeurs manquantes par la moyenne ou la médiane de toutes les valeurs de l'attribut ou des valeurs appartenant à la même classe. Des méthodes plus sophistiquées estiment la valeur la plus probable par régression, inférence bayésienne ou arbre de décision. L'étape de nettoyage peut aussi éliminer les données bruitées et valeurs aberrantes. Leur détection s'effectue par le calcul de l'écart interquartile ou par des méthodes de groupement des données par classe (*binning*), régression et *clustering*.

1.2.3.3 Réduction de la dimension

L'objectif de la réduction de dimension est l'obtention d'une représentation réduite du jeu de données tout en minimisant la perte d'information et conservant l'intégrité des données. L'analyse d'un ensemble de donnée réduit est plus performante, compréhensible et rapide. Les méthodes de réduction cherchent à découvrir les attributs non pertinents et la corrélation entre les attributs. Il existe deux approches pour la réduction de dimension : la sélection d'attributs (*feature selection*) qui élimine les attributs non pertinents et corrélés et la redescription d'attributs en appliquant des transformations (exemple : analyse en composantes principales).

1.2.3.4 Transformation et discrétisation

Les méthodes de transformation convertissent les valeurs des attributs afin d'améliorer la performance d'algorithmes de classification tels que ceux basés sur une distance. La transformation peut se faire avec une fonction mathématique simple (puissance, racine carrée, logarithme décimal, inverse, etc.) ou par la standardisation (centrer-réduire). Cette dernière procure un poids égal et une caractéristique particulière pour l'ensemble d'attributs. La standardisation peut s'effectuer par une méthode **min-max** qui remet à l'échelle les valeurs d'un attribut A dans un nouvel intervalle spécifié $[new\_min_A, new\_max_A]$:

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \tag{1.4}$$

Une autre méthode est la standardisation **z-score** selon la formule

$$v'_i = \frac{v_i - \mu_A}{\sigma_A} \tag{1.5}$$

où μ_A et σ_A sont la moyenne et l'écart type de la variable A respectivement. La nouvelle variable aura une moyenne égale à 0 et un écart type égale à 1.
Plusieurs algorithmes d'apprentissage ne peuvent pas traiter les valeurs continues. Ainsi, il est nécessaire de transformer les attributs numériques continues en attributs numériques discrets ou ordinaux (ce principe se nomme la discrétisation). Les techniques de discrétisation incluent le *binning*, le *clustering*, l'analyse d'histogrammes, les arbres de décision et l'analyse de corrélation.

1.3 Apprentissage automatique

L'apprentissage automatique (*machine learning* en anglais) s'intéresse à la conception, le développement et l'application d'algorithmes qui apprennent et évoluent avec des expériences (Mitchell, 1997) pour découvrir des connaissances (interprétation et description) ou faire des décisions (prédictions). Les expériences, appelées

concepts, sont un ensemble d'exemples caractérisés par des attributs. Ces algorithmes d'apprentissage produisent des modèles (*descriptions du concept*) qui décrivent les relations qui existent entre l'ensemble d'attributs. Ces modèles peuvent être exploités et appliqués sur des exemples futurs.

Les approches d'apprentissage automatique se divisent en trois catégories selon le type de la décision produite. La première catégorie est la **classification**. C'est la tâche d'assigner des objets inconnus à des catégories (classes ou étiquettes) prédéfinies en utilisant un modèle construit à partir d'un ensemble d'exemples connus et étiquetés. La deuxième catégorie est la **prédiction numérique**. Contrairement à la classification où la décision est catégorielle (discrète et non ordonnée), la prédiction numérique produit une décision numérique continue. Elle peut se faire par des méthodes d'analyse de régression. La classification et la prédiction numérique sont des approches dites *supervisées* à cause de la connaissance préalable des classes des exemples. En outre, il y a l'apprentissage *non supervisé* où les objets ne sont pas étiquetés. Ça correspond à la troisième catégories qui est le **clustering** (catégorisation). Son objectif est de découvrir les regroupements et les partitions naturels d'un ensemble d'objets. En analysant la forme, la taille et la densité des groupes, les algorithmes cherchent à maximiser les similarités entre les objets du même groupe et à minimiser les similarités entre les objets des groupes différents.

Les sous-sections suivantes détaillent quelques approches principales d'apprentissage supervisé et non supervisé qui ont été utilisées ou évaluées dans les études décrites dans ce mémoire.

1.3.1 Approches supervisées

1.3.1.1 Apprentissage par arbres de décision

L'apprentissage basé sur des arbres de décision utilise une structure d'arbre intuitive qui facilite la description et l'interprétation du modèle d'apprentissage (figure 1.3). Dans cette structure, chaque nœud interne correspond à un test sur un attribut, chaque branche représente le résultat du test et chaque feuille contient une classe (chaque classe peut correspondre plusieurs feuilles).

La classification des objets se fait par une séquence de tests successifs sur les attributs qui les décrivent. Il existe un nombre exponentiel d'arbres de décision construits à partir d'un ensemble d'attributs. L'algorithme Hunt (Hunt *et al.*, 1966) adopte une stratégie récursive basée sur le principe diviser-pour-régner pour construire un arbre optimal (voir algorithme 1). Cet algorithme est à la base d'un ensemble d'algorithmes d'arbres de décision célèbres tels que ID3 (Quinlan, 1979), C4.5 (Quinlan, 1993) et CART (Breiman *et al.*, 1984).

Algorithme 1 : Construction d'un arbre de décision (Cornuéjols et Miclet, 2010)

Procédure `Construire-arbre(`*nœud* X`)`

début

si *Tous les objets de* X *appartiennent à la même classe* **alors**
Créer une feuille portant le nom de cette classe

sinon

Sélectionner le **meilleur** attribut pour créer un nœud
Le test associé à ce nœud sépare X en deux parties notées X_g et X_d
`Construire-arbre(`*nœud* X_g`)`
`Construire-arbre(`*nœud* X_d`)`

Dans chaque récursion, l'algorithme sélectionne le meilleur attribut pour parti-

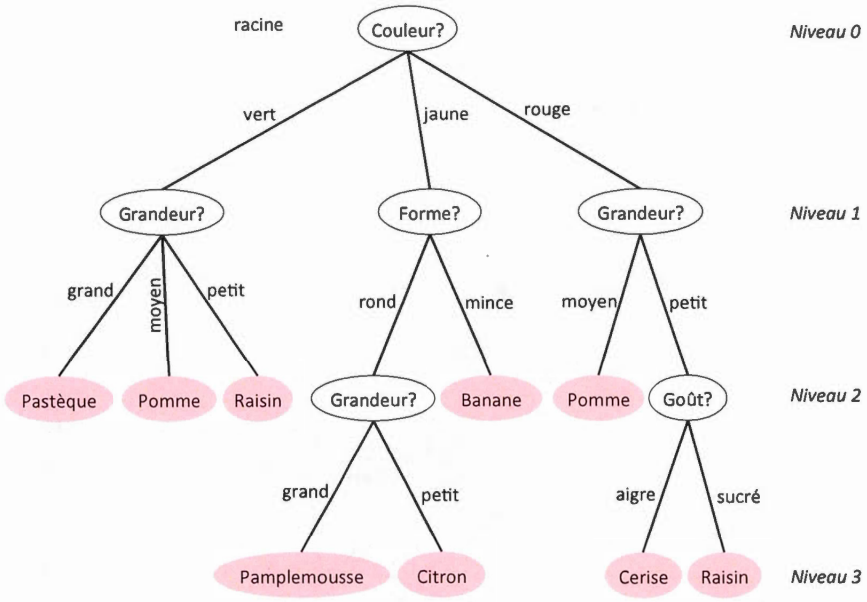


Figure 1.3: Un exemple d'arbre de décision. Adaptée de Duda *et al.* (2001).

tionner un groupe (nœud) en sous-ensembles purs (un groupe pur contient des objets appartenant à la même classe). Il existe plusieurs mesures pour classer et sélectionner le meilleur attribut, parmi eux le *gain d'information* (utilisé par ID3) et l'*indice de Gini* (utilisé par CART). Supposons qu'on a un nœud N contenant un ensemble d'objets D étiquetés par m classes, C_i ($pour i = 1, ..., m$). $C_{i,D}$ est l'ensemble d'objets de la classe C_i dans D. La probabilité qu'un objet de classe C_i appartient à D (p_i) est estimée par $|C_{i,D}|/|D|$.

Le gain d'information pour un attribut A est donné par la formule 1.6.

$$Gain(D, A) = Entropie(D) - \sum_{v \in valeur(A)} \frac{|D_v|}{|D|} \times Entropie(D_v), \qquad (1.6)$$

où D_v est le sous ensemble de D qui contient la valeur v de l'attribut A et $Entropie(D)$ est l'*entropie de Shannon* (quantité d'information) calculée par la

formule 1.7 (Cover et Thomas, 2006).

$$Entropie(D) = -\sum_{i=1}^{m} p_i log_2(p_i). \tag{1.7}$$

L'attribut A avec le gain le plus grand est choisi pour diviser le nœud N. L'algorithme C4.5 utilise une version modifiée de gain d'information appelée *ratio du gain d'information.*

L'indice de *Gini* mesure l'impureté d'un ensemble d'objets. Si on divise l'ensemble D du nœud N en v sous-ensembles $\{D_1, D_2, ..., D_v\}$ selon les valeurs de l'attribut A, l'indice de *Gini* de A sera :

$$Gini(A) = \sum_{i=1}^{v} \frac{|D_v|}{|D|} \times Gini(D_v), \tag{1.8}$$

où $Gini(D_v)$ est l'indice de *Gini* du sous-ensemble D_v donné par la formule :

$$Gini(D_v) = 1 - \sum_{i=1}^{m} p_i^2. \tag{1.9}$$

L'attribut qui maximise la réduction d'impureté (qui a l'indice de *Gini* le plus petit) sera sélectionné pour la division.

1.3.1.2 Approches bayésiennes

Les approches bayésiennes calculent les probabilités d'appartenance d'un objet donné à une classe particulière. Elles se basent sur le **théorème de Bayes** qui a la formule suivante :

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{1.10}$$

X et Y sont deux variables aléatoires. $P(Y|X)$ est la probabilité *a posteriori* de Y sachant X. $P(Y)$ et $P(X)$ sont les probabilités *a priori* de Y et X respectivement. $P(X|Y)$ est la probabilité *a posteriori* de X sachant Y.

Le théorème de Bayes peut être utilisé pour résoudre des problèmes de classification. Si un objet $\mathbf{X}$ (décrit par un ensemble de d attributs) et une classe variable

Y ont une relation non déterministe, ils seront considérés comme deux variables aléatoires. À partir des données d'apprentissage, on peut estimer les probabilités *a posteriori* $P(Y|\mathbf{X})$ de chaque couple $\mathbf{X}$ et Y. La connaissance de ces probabilités permet la classification d'un nouvel objet $\mathbf{X}'$ par l'identification de la classe Y' qui maximise la probabilité *a posteriori* $P(Y'|\mathbf{X}')$ (Tan *et al.*, 2006).

La *classification bayésienne naïve* suppose que les attributs sont conditionnellement indépendants étant donnée une classe y. L'indépendance des attributs est exprimée par la formule :

$$P(\mathbf{X}|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y), \tag{1.11}$$

Pour classifier un objet, on calcule la probabilité *a posteriori* de chaque classe Y selon :

$$P(Y|\mathbf{X}) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(\mathbf{X})} \tag{1.12}$$

Étant donnée la probabilité *a priori* $P(\mathbf{X})$ est constante pour toute les classes Y, le prédicteur choisit la classe qui maximise le terme $\prod_{i=1}^{d} P(X_i|Y)$ (Tan *et al.*, 2006).

Contrairement à l'approche naïve, les *réseaux bayésiens* relâchent l'hypothèse d'indépendance entre les variables et permettent des dépendances entre les sous-ensembles de variables. Ils représentent les relations probabilistes entre les variables par un modèle graphique. Un réseau bayésien entrainé sur un ensemble de variables aléatoires est constitué par un graphe orienté acyclique et des tables de probabilités conditionnelles. Les sommets du graphe représentent des variables et les arêtes décrivent les relations de dépendance entre deux sommets (parent et descendant). Chaque sommet est associé à une table de probabilité conditionnelle qui spécifie la distribution conditionnelle du sommet sachant ses parents. Le réseau bayésien respecte la propriété de l'*indépendance conditionnelle* où chaque

sommet est conditionnellement indépendant de ses non-descendants étant donnés ses parents (Duda *et al.*, 2001; Tan *et al.*, 2006; Han *et al.*, 2011).

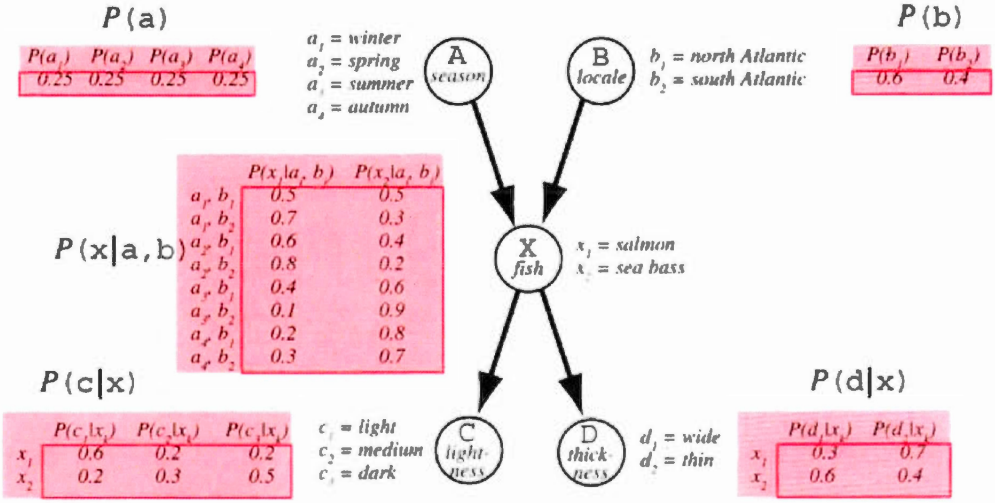


Figure 1.4: Réseau bayésien. Source : Duda *et al.* (2001).

1.3.1.3 Séparateurs à vaste marge

Les séparateurs à vaste marge sont connus aussi par le nom *machines à vecteurs de support* ou en anglais *support vector machine* (*SVM*). Dans un problème de classification binaire, ces approches séparent les exemples (objets) en deux classes par un hyperplan en utilisant des exemples essentiels appelés *vecteurs de support* et des marges définies par ces derniers (figure 1.5) (Boser *et al.*, 1992). Il existe un nombre infini d'hyperplans séparateurs qui ont un taux d'erreur de classification d'exemples d'entrainement nul. Néanmoins, ces hyperplans ne garantissent pas une performance égale avec des exemples inconnus. Pour cela, les SVM cherchent à trouver l'hyperplan qui minimise le risque empirique de classification (le nombre d'exemples de test mal classés). Selon un principe de statistique d'apprentissage (Vapnik et Chervonenkis, 1971), Boser *et al.* (1992) ont montré que l'hyperplan

avec une marge maximale minimise ce risque.

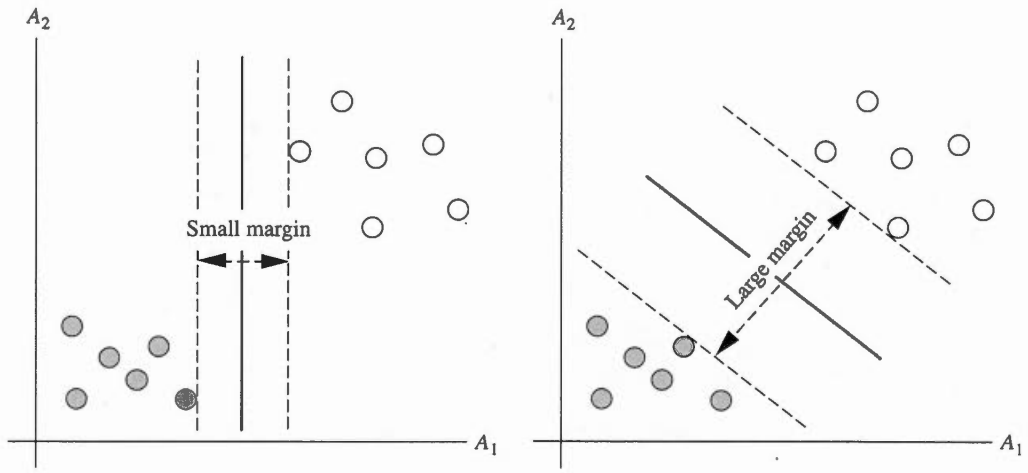


Figure 1.5: Séparations linéaires des données (Han *et al.*, 2011).

Supposons qu'on a un ensemble de N objets notés par $(\mathbf{x_i}, y_i)$ $(i = 1, 2, ..., N)$ où $\mathbf{x_i}$ est l'ensemble d'attributs du $i^{ème}$ exemple associé avec la classe y_i et $y_i \in \{-1, 1\}$. Selon la propriété de séparabilité linéaire des exemples, Il existe deux types de SVM :

— *SVM linéaires* appliqués sur les données linéairement séparables. L'hyperplan séparateur est décrit par l'équation linéaire suivante :

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \tag{1.13}$$

où $\mathbf{w}$ est un vecteur de poids (même taille que $\mathbf{x}$) et b est un coefficient. À partir de l'équation 1.13, on peut déduire la classe de chaque exemple $\mathbf{z}$ selon le signe du terme $\mathbf{w} \cdot \mathbf{z} + b$, s'il est positif $y = 1$ ($\mathbf{z}$ se trouve en haut du séparateur), sinon $y = -1$ ($\mathbf{z}$ en bas du séparateur).

La marge de l'hyperplan séparateur est la distance entre deux hyperplans

parallèles définis par les équations 1.14 et 1.15.

$$\mathbf{w} \cdot \mathbf{x} + b = 1, \tag{1.14}$$

$$\mathbf{w} \cdot \mathbf{x} + b = -1. \tag{1.15}$$

Par soustraction de l'équation 1.15 de 1.14, on peut calculer la distance entre ces deux hyperplans, qui est :

$$d = \frac{2}{||\mathbf{w}||}. \tag{1.16}$$

où $||\mathbf{w}||$ est la norme du vecteur des poids $\mathbf{w}$.
La recherche de la marge maximale revient à : (1) l'estimation des paramètres $\mathbf{w}$ et b du séparateur à partir des données d'entrainement selon ces deux conditions :

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq 1 \quad \text{si} \quad y_i = 1 \quad \text{et} \tag{1.17}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1 \quad \text{si} \quad y_i = -1 \tag{1.18}$$

qui peuvent être résumées par :

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \quad i = 1, 2, ..., N. \tag{1.19}$$

et de (2) la minimisation de la fonction objective :

$$f(\mathbf{w}) = \frac{||\mathbf{w}||^2}{2}. \tag{1.20}$$

La minimisation de la fonction quadratique (1.20) sous la contrainte linéaire (1.19) est un problème d'optimisation quadratique convexe. Ce problème peut être résolu par la réécriture de la fonction et de la contrainte en une fonction *lagrangienne* primaire :

$$L_p = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{N} \alpha_i \Big(y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1 \Big), \tag{1.21}$$

où α_i sont des coefficients positifs appelés *multiplicateurs de Lagrange*, puis sa transformation en une formulation duale (avec les conditions de **Kuhn-Tucker**) qui implique seulement les multiplicateurs de Lagrange et les données d'entrainement :

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}. \tag{1.22}$$

La résolution du problème dual peut s'effectuer par des techniques d'optimisation quadratique.

Dans l'étape de test, la classe $y_\mathbf{z}$ d'un exemple inconnu $\mathbf{z}$ est donnée par une fonction basée sur la formulation lagrangienne :

$$f(\mathbf{z}) = signe(\mathbf{w} \cdot \mathbf{z} + b) = signe\Big(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i} \cdot \mathbf{z} + b \Big). \tag{1.23}$$

Si $f(\mathbf{z})$ est positive, $y_\mathbf{z} = 1$, sinon $y_\mathbf{z} = -1$ (Tan *et al.*, 2006; Cornuéjols et Miclet, 2010; Han *et al.*, 2011).

— *SVM non linéaires* pour les données non séparables linéairement. Ce sont une extension des SVM linéaires. Ils transforment l'espace des données originales en un espace de redescription de dimension supérieure (en utilisant une fonction de mappage $\Phi(\mathbf{x})$) afin de trouver un hyperplan séparateur linéaire dans le nouvel espace. L'équation duale 1.22 devient :

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j}). \tag{1.24}$$

Le produit scalaire $\Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$ correspond à une mesure de similarité entre les deux vecteurs dans l'espace de redescription. Ce produit peut être remplacé par une fonction de similarité $K(\mathbf{x_i}, \mathbf{x_j})$, appelée **fonction de noyau** et qui se fait dans l'espace original. Parmi les fonctions de noyau admissibles, les SVM non linéaires utilisent des fonctions polynomiales, des fonctions gaussiennes à base radiale (radial basis function, *RBF*) et des fonctions sigmoïdes (Han *et al.*, 2011; Tan *et al.*, 2006).

1.3.1.4 Les k plus proches voisins

La méthode des k plus proches voisins (*k-nearest neighbor* ou *KNN* en anglais) appartient à la famille des approches d'apprentissage qui se base sur les exemples. Contrairement aux approches vues précédemment (arbres de décision, bayésiens, etc.), celles-là ne construisent pas un modèle ou une hypothèse d'entrainement. Elles conservent les objets d'entrainement pour les utiliser ultérieurement dans la classification des objets inconnus. La classification d'un objet donné par la méthode des k plus proches voisins s'effectue par la recherche et la comparaison avec les k objets d'entrainement proches (similaires) à lui. La proximité entre les objets est définie par une distance telle que la distance euclidienne. La distance euclidienne entre deux objets X_1 et X_2 décrits respectivement par les deux ensembles d'attributs $(x_{11}, x_{12}, ..., x_{1n})$ et $(x_{21}, x_{22}, ..., x_{2n})$ est :

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}. \tag{1.25}$$

D'autres types de distances peuvent être appliqués tels que la distance de Hamming pour les valeurs discrètes. Une fois que la liste des objets proches est déterminée, l'objet inconnu est assigné à la classe majoritaire de la liste.

1.3.1.5 Approches ensemblistes

Ces approches tendent à améliorer la performance de la classification par une combinaison de plusieurs modèles de base ayant des performances faibles sur un jeu d'entrainement. La prédiction finale d'un exemple de test est constituée par le vote majoritaire de ces modèles de base. Néanmoins il faut que les modèles de bases soient indépendants de l'un de l'autre et aient un comportement différent de l'aléatoire (taux d'erreur inférieur à 0.5) pour que leur combinaison aie une meilleure performance qu'un modèle de base. Si ϵ_i est le taux d'erreur du modèle

i, le taux d'erreur de l'ensemble des modèles de base est :

$$\epsilon_{ensemble} = \sum_{i=\frac{N}{2}}^{N} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \tag{1.26}$$

On trouve dans la littérature plusieurs méthodes pour la construction d'un modèle ensembliste (Tan *et al.*, 2006; Witten *et al.*, 2011). Ces méthodes peuvent manipuler les jeux de données, les attributs, les classes et/ou les algorithmes d'entrainement.

— *Génération de plusieurs jeux d'entrainement par une série d'échantillonnages du jeu de données initial.* Les échantillonnages se font selon une distribution donnée et ont la même taille. Les modèles de base sont entrainés avec ces nouveaux jeux de données afin de constituer le modèle ensembliste. L'algorithme **Bagging** (de *bootstrap aggregating*) est basé sur cette méthode. Chaque échantillonnage (*bootstrap*) se fait avec remise, par conséquent des exemples peuvent être présents plusieurs fois dans le même *bootstrap* comme ils peuvent ne pas y apparaitre. L'algorithme **Boosting** et sa variante **AdaBoost** (de *Adaptive Boosting*) attribuent des poids aux exemples dans chaque *bootstrap*. Initialement, tous les exemples du premier *bootstrap* ont le même poids. Un modèle de base est entrainé sur ce jeu de données et les poids des exemples seront mis à jour (augmentés si les exemples sont mal classifiés sinon réduits). Cette étape est répétée plusieurs fois en sélectionnant des exemples qui ont un poids plus élevé (se concentrant sur les cas difficiles à classifier).

— *Sélection aléatoire d'un sous ensemble d'attributs pendant la génération des jeux d'entrainement.* L'algorithme **Random forest** est une combinaison de ce principe avec l'algorithme *bagging* en utilisant des arbres de décision comme modèles de base. Dans chaque itération, *Random forest* entraine un arbre sur un *bootstrap* (échantillonnage aléatoire avec remise) en sélec-

tionnant d'une manière aléatoire un ensemble d'attributs pour la division des nœuds de l'arbre.

— *Transformation du jeu de données multi-classe en un problème de classification binaire par partitionnement aléatoire de l'ensemble des classes en deux sous-ensembles disjoints correspondant à deux classes.* Les classes des exemples appartenant à chaque sous ensemble sont réétiquetées. La transformation est répétée plusieurs fois avec une construction d'un modèle de base bi-classe. Un system de vote est utilisé pour déterminer la classe d'un exemple de test à partir des prédictions des modèles de base.

— *Manipulation des algorithmes d'apprentissage.* Le modèle final peut être homogène ou hétérogène. Un modèle ensembliste est homogène quand tous ces modèles de base appartiennent à la même classe d'algorithmes (arbre de décision ou SVM, etc.). Les modèles de base sont entrainés sur le même jeu de données par contre paramétrés différemment. Par exemple attribuer des poids différents dans chaque itération ou choisir un attribut parmi un top-k pour la division des nœuds dans un arbre au lieu de chercher le meilleur attribut. Les modèles hétérogènes combinent plusieurs algorithmes de types différents (par exemple un arbre de décision, un bayésien et un *KNN*). Cette approche est appelée **Stacked generalization** ou **Stacking**. Au lieu d'utiliser le concept de vote pour la classification, le *stacking* utilise un méta-modèle qui apprend quels modèles de base classifient correctement les données.

1.3.2 Approches non supervisées

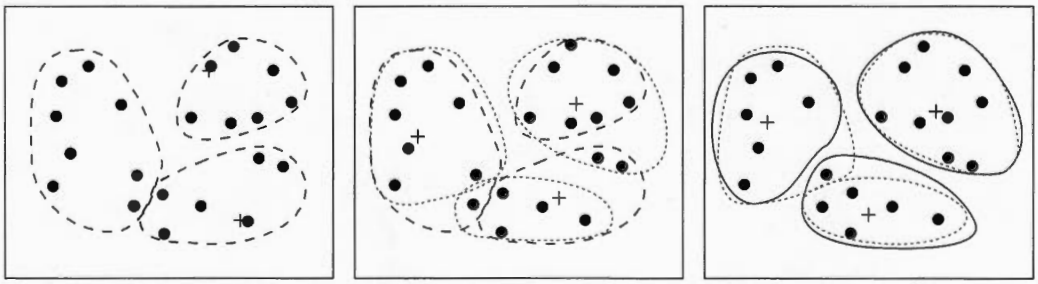
1.3.2.1 K-moyennes et k-médoïdes

Les algorithmes des k-moyennes (k-*means*) et k-médoïdes cherchent à trouver une partition de k groupes exclusives qui optimise un critère de partitionnement (fi-

gure 1.6). Chaque groupe (*cluster*) sera représenté par son centre (soit la moyenne du groupe, ne faisant pas obligatoirement partie de l'ensemble dans le cas de k-moyennes, soit un point médiane du groupe, appartenant à l'ensemble des données dans le cas de k-médoïdes). Ces approches tendent à minimiser la distance (exprimée par l'*erreur quadratique*) entre chaque objet et le centre de son groupe afin de maximiser la similarité des objets à l'intérieur d'un *cluster*. Cela revient à minimiser la fonction objective *somme des erreurs quadratiques* de tous les objets avec leurs centres ; sa formule est :

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2, \tag{1.27}$$

où k est le nombre de groupes, p est un objet, c_i est le centre du cluster C_i et $dist(x, y)$ est la distance entre les deux points x et y qui peut être calculée par la formule 1.25 (Han *et al.*, 2011). Ce problème d'optimisation est NP-difficile même pour $k = 2$ (Jain, 2010).



(a) Initial clustering **(b)** Iterate **(c)** Final clustering

Figure 1.6: Partitionnement d'un ensemble d'objets par la méthode k-moyennes. Source : Han *et al.* (2011).

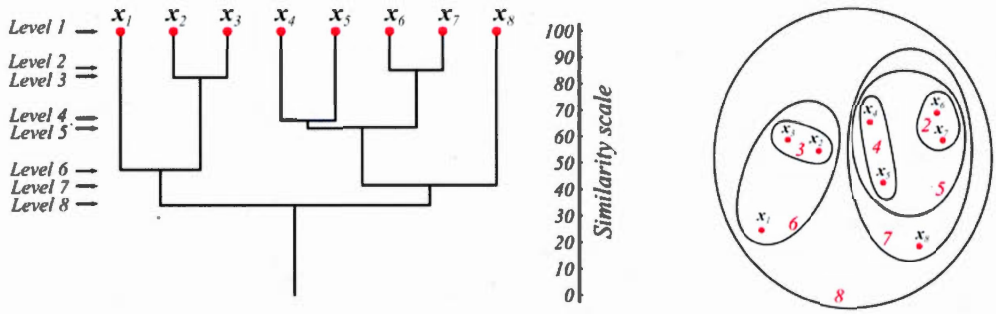
1.3.2.2 Regroupement hiérarchique

Les méthodes de regroupement hiérarchique produisent une hiérarchie de partitions (contrairement aux méthodes de partitionnement qui produisent des parti-

tions disjointes à un seul niveau). Cette hiérarchie peut être représentée par un arbre appelé dendrogramme, par un diagramme de Venn ou textuellement (figure 1.7). Han *et al.* (2011) divise les méthodes hiérarchiques en trois catégories : méthodes *algorithmiques*, méthodes *probabilistes* et méthodes *bayésiennes*. Les méthodes *algorithmiques* considèrent l'ensemble des objets comme déterministe et génèrent les partitions selon une distance déterministe entre les objets. Cependant, les méthodes *probabilistes* utilisent des modèles probabilistes pour calculer la distance entre les objets et la qualité des clusters. Les méthodes *bayésiennes* calculent une distribution de l'ensemble des partitions possibles. Historiquement, les méthodes *algorithmiques* sont les plus utilisées. Elles comprennent deux approches basiques : une approche ascendante (agglomérative) et une approche descendante. À l'initialisation, l'approche ascendante considère les objets comme des clusters individuels puis, à chaque étape, elle fusionne les deux paires de clusters les plus proches. L'autre approche commence par l'ensemble des objets comme un seul cluster puis, à chaque itération, elle sélectionne et divise un cluster jusqu'à ce que les clusters contiennent un seul objet.

1.3.2.3 Algorithmes basés sur la densité

Les algorithmes basés sur la densité définissent un cluster comme une région de grande densité d'objets entourée par une région de faible densité. DBSCAN (*Density-Based Spatial Clustering of Applications with Noise* (Ester *et al.*, 1996)) est le plus connu parmi ces algorithmes. Les objets dans les régions de faible densité sont considérés comme du bruit et seront éliminés. La densité pour un objet donné est calculée par le nombre d'objets qui tombent dans un rayon *Eps* à partir de cet objet. Afin de déterminer si le voisinage est dense ou non, l'algorithme utilise un autre paramètre $MinPts$ qui spécifie le seuil de densité des régions. Cette approche de densité permet de classifier chaque objet comme 1) un objet central qui existe à l'intérieur d'une région dense, 2) un objet de bordure qui est très



a) Dendrogramme

b) Diagramme de Venn

$$\{\{\mathbf{x}_1, \{\mathbf{x}_2, \mathbf{x}_3\}\}, \{\{\{\mathbf{x}_4, \mathbf{x}_5\}, \{\mathbf{x}_6, \mathbf{x}_7\}\}, \mathbf{x}_8\}\}$$

c) Représentation textuelle

Figure 1.7: Représentations graphiques des résultats d'algorithmes hiérarchiques. Adaptée de Duda *et al.* (2001).

proche d'une ou plusieurs régions denses ou 3) un objet bruit qui existe dans une région de faible densité. Un cluster est constitué par l'ensemble des objets centraux qui se trouvent mutuellement dans les voisinages et par leurs objets de bordure. Les approches basées sur la densité sont utilisées pour extraire des clusters irréguliers, arbitraires ou entrelacés ; à l'opposé des approches de partitionnement et hiérarchiques qui sont conçues pour trouver des clusters de formes sphériques.

1.3.3 Évaluation de l'apprentissage

Un modèle de classification est robuste s'il peut classifier correctement un jeu de données indépendant des objets d'entrainement. Cela se traduit par le pouvoir de *généralisation* du modèle. Usuellement, la performance d'un modèle est évaluée par son **taux d'erreur** sur un jeu de *test*. C'est la proportion des exemples faussement classifiés. À l'inverse, la proportion des exemples correctement classifiés

est le **taux de bonne classification** ou *accuracy*. Cependant, ces mesures considèrent chaque classe d'importance égale, ce qui n'est pas toujours vrai dans les problèmes de classification. Dans les cas de déséquilibre de taille des classes (*class imbalance problem*) on peut se fier à des métriques plus adéquates et fiables pour mesurer l'évaluation des modèles d'apprentissage.

1.3.3.1 Métriques d'évaluation de la performance

Supposons qu'un jeu de données divisé par deux classes : une classe d'intérêt et une classe négative étiquetées + et − respectivement. Les prédictions d'un modèle testé sur ce jeu de données peuvent être :

- Correctes pour la classe + (ensemble des vrais positifs)
- Correctes pour la classe − (ensemble des vrais négatifs)
- Fausses pour la classe −, prédites comme + (ensemble des faux positifs)
- Fausses pour la classe +, prédites comme − (ensemble des faux négatifs)

Ces quatre ensembles sont agencés dans une **matrice de confusion** (tableau 1.2).

Tableau 1.2: Matrice de confusion pour un problème de classification binaire

		Classes prédites		
		+	−	Total
Classes réelles	+	Vrais positifs (VP)	Faux négatifs (FN)	P
	−	Faux positifs (FP)	Vrais négatifs (VN)	N
	Total	P'	N'	P + N

Si on généralise le problème à n classes $(C_1, C_2, ..., C_n)$, les résultats de prédiction de la classe C_i $(1 \leq i \leq n)$ seront :

- VP de C_i est l'ensemble de tous les exemples C_i classés comme C_i,
- VN de C_i est l'ensemble de tous les exemples différents de C_i classés comme

non C_i,

— FP de C_i est l'ensemble de tous les exemples différents de C_i classés comme C_i,

— FN de C_i est l'ensemble de tous les exemples C_i classés comme non C_i.

À partir de la matrice de confusion, plusieurs métriques peuvent être calculées :

1. **Taux de vrais positifs (TVP)** : proportion des exemples positifs correctement prédits. Il est appelé aussi *sensibilité*, *rappel* ou *TPR* (pour *true positive rate).*

$$TVP = VP/(VP + FN) \tag{1.28}$$

2. **Taux de vrais négatifs (TVN)** : proportion des exemples négatifs correctement prédits. Il est connu aussi par la *spécificité.*

$$TVN = VN/(VN + FP) \tag{1.29}$$

3. **Taux de faux positifs (TFP)** : proportion des exemples négatifs incorrectement classés. On l'appelle aussi *FPR* (pour *false positive rate*).

$$TFP = FP/(VN + FP) \tag{1.30}$$

4. **Taux de faux négatifs (TFN)** : proportion des exemples positifs incorrectement classées.

$$TFN = FN/(VP + FN) \tag{1.31}$$

5. **Précision** : fraction des exemples positifs correctement classés par rapport à tous les exemples classés positifs par le modèle.

$$Précision = VP/(VP + FP) \tag{1.32}$$

6. **F-mesure** : moyenne harmonique du rappel et de la précision.

$$F - mesure = \frac{2 \times rappel \times précision}{rappel + précision} \tag{1.33}$$

1.3.3.2 Méthodes d'estimation du risque réel

— **Utilisation d'un échantillon de test (méthode *holdout*)** : le jeu de données initial est divisé en deux sous-ensembles disjoints. L'un des sous-ensembles est utilisé pour l'entrainement d'un modèle d'apprentissage (*training set*) et l'autre est utilisé pour évaluer la performance du modèle (*test set*). Le *risque empirique* est l'ensemble des erreurs commises sur le jeu d'entrainement. Par contre, la mesure des erreurs commises sur le jeu de test est considérée comme une estimation du *risque réel*. Une variante de cette méthode est le **sous-échantillonnage aléatoire**. La méthode *holdout* est répétée k fois sur un jeu de données. Le taux d'erreur global est la moyenne des erreurs de chaque itération.

— **Validation croisée (*k-fold cross-validation*)** : cette méthode divise le jeu de données de taille d en k partitions disjointes et de tailles approximativement égales ($2 \leq k \leq d$). L'entrainement et le test sont réalisés k fois alternativement. Dans chaque itération, une partition est gardée pour le test et le reste des *k-1* partitions est fusionné pour construire un modèle d'apprentissage. ***Leave-one-out*** est un cas spécial de la validation croisée où $k = d$. Le taux d'erreur global est la moyenne des erreurs de toutes les itérations.

— **Bootstrap** : la méthode réalise un échantillonnage avec remise du jeu de données original. Si ce dernier contient d exemples, un échantillon (ou *bootstrap*) de taille d contiendrait 63.2% des exemples des données originales (comme un exemple peut apparaitre plusieurs fois, sa probabilité d'être sélectionnée est $1-(1-1/d)^d$ et elle tend vers $1-e^{-1} = 0.632$ quand d est assez grand). Le reste des exemples non sélectionnés (36.8%) forme le jeu de test. L'échantillonnage est répété k fois. Dans chaque itération i, un modèle est entrainé avec le bootstrap $boot_i$ et testé sur le jeu de test $test_i$

et sur la totalité du jeu original. La méthode **.632 bootstrap** calcule le taux de bonne classification (*accuracy*) comme suit : .

$$Accuracy = \frac{1}{k}\sum_{i=1}^{k}(0.632 \times accuracy_{test_i} + 0.368 \times accuracy_{original}) \quad (1.34)$$

1.4 Apprentissage automatique et bioinformatique

L'apprentissage automatique et la biologie ont une association étroite et complexe. Les travaux pionniers d'apprentissage automatique ont essayé de mimer des concepts de la biologie tels que les neurones (perceptrons et réseaux de neurones artificiels (Rosenblatt, 1958; Rumelhart *et al.*, 1986)) et l'évolution (algorithmes génétiques (Holland, 1992)). La biologie a appliqué aussi des concepts d'apprentissage automatique pour la modélisation de ses problématiques. Les premières applications des techniques d'apprentissage automatique dans les sciences biomédicales remontent aux années 1970 (figure 1.8). Les premiers algorithmes utilisés étaient les réseaux de neurones artificiels (*ANN*), les modèles de Markov cachés (*HMM*) et les arbres de décisions. Les autres algorithmes, tels que les *SVM* et *Random Forest*, ont commencé à être appliqués en bioinformatique seulement à la fin des années 1990 (figure 1.8).

L'apprentissage supervisé a été utilisé dans différents domaines de la bioinformatique pour la classification des données biologiques et génomiques. Des algorithmes ont été conçus pour l'identification et la prédiction de nombreux éléments associés aux gènes tels que les promoteurs (Bucher, 1990), les sites d'épissages (Degroeve *et al.*, 2002), les sites d'initiation à la transcription (Ohler *et al.*, 2002) et les activateurs (Heintzman *et al.*, 2007). D'autres approches d'annotation des gènes combinent des modèles entrainés pour prédire les types des régions des gènes comme les introns, exons et les régions non traduites (*UTR*) (Picardi et Pesole, 2010).

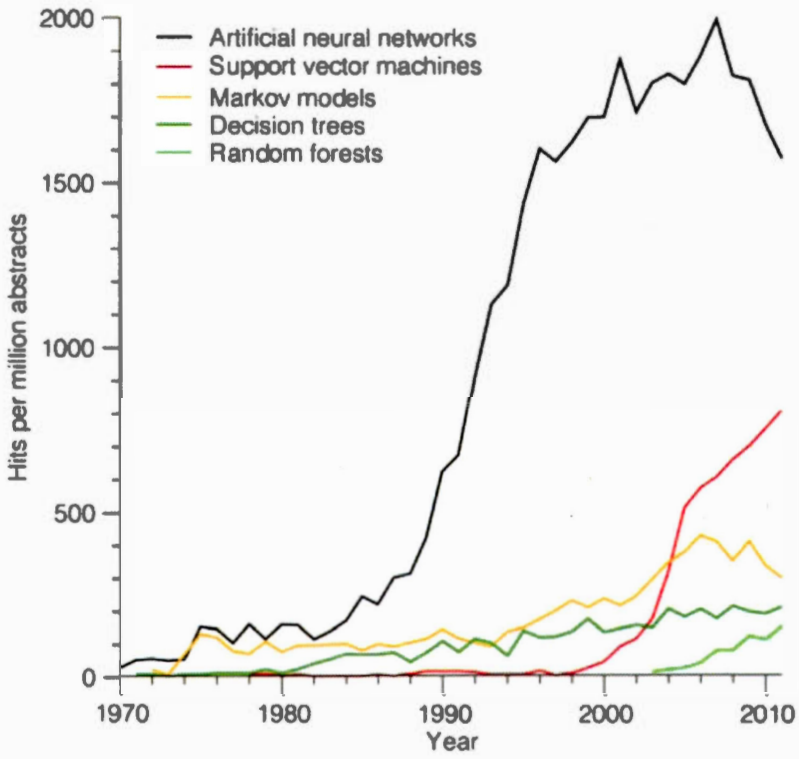


Figure 1.8: La croissance de l'utilisation des méthodes d'apprentissage supervisé dans les articles référencés dans PubMed. Source : Jensen et Bateman (2011).

Les approches d'apprentissage supervisé ont été largement exploitées dans la prédiction de nouveaux types d'ARNs non codants à partir des données de séquençage des ARNs (*RNA-seq*). Différents outils, basés sur ces approches, ont été proposés pour la prédiction des miARNs et leurs précurseurs dans plusieurs espèces. Par exemple, triplet-SVM (Xue *et al.*, 2005) et MiRFinder (Huang *et al.*, 2007) entrainent des modèles *SVM* pour prédire les précurseurs des miARNs. D'autres outils implémentent des modèles basés sur les *HMM* comme ProMiR (Nam *et al.*, 2005), HHMMiR (Kadri *et al.*, 2009) et SSCprofiler (Oulas *et al.*, 2009). Des modèles probabilistes basés sur des approches bayésiennes ont été développés pour l'identification des séquence matures des miARNs (BayesMiRNAfind (Yousef

et al., 2006), MatureBayes (Gkirtzou *et al.*, 2010)). MiPred (Jiang *et al.*, 2007), HuntMi (Gudyś *et al.*, 2013) utilisent *Random forest* pour prédire les précurseurs des miARNs et MiRdup (Leclercq *et al.*, 2013) l'utilise pour valider les miARNs matures.

Les méthodes d'apprentissage automatique peuvent servir à faire du génotypage, de la classification fonctionnelle et de l'annotation de séquences. Ces méthodes sont spécifiques soit à une espèce soit à un groupe d'organismes donné. Particulièrement chez les virus, il existe peu de méthodes de classification qui intègrent des techniques d'apprentissage automatique. SCUEAL (Pond *et al.*, 2009), MuLDAS (Kim *et al.*, 2010) et COMET (Struck *et al.*, 2014) sont des outils conçus pour la classification des virus VIH-1 ainsi d'autres virus (VHC). SCUEAL combine un modèle phylogénétique et un algorithme génétique alors que COMET est basé sur des modèles de Markov et des arbres de décision. Quant à MuLDAS, il entraine un modèle statistique basé sur l'analyse discriminante linéaire.

Plusieurs méthodes basées sur l'apprentissage automatique sont développées pour la classification des séquences bactériennes, métagénomiques et écologiques en rangs taxonomiques. Elle sont compétitives avec des méthodes basées sur l'alignement de séquences (Altschul *et al.*, 1990) et la phylogénie (Matsen *et al.*, 2010). Les deux outils Ribosomal Database Project Classifier (RDP) (Wang *et al.*, 2007) et Naive Bayes Classifier (NBC) (Rosen *et al.*, 2011) implémentent un algorithme bayésien naïf pour entrainer leurs modèles de classification. MgFC (MetaGenomic Fragment Classification) (Tzahor *et al.*, 2009) et PhyloPythiaS (Patil *et al.*, 2011) utilisent les séparateurs à vaste marge (*SVM*). Les modèles de Markov d'ordre variable (ou des modèles interpolés, en anglais Interpolated Markov Models (*IMM*)) ont été utilisés initialement dans l'identification des gènes dans des génomes microbiens (GLIMMER (Salzberg *et al.*, 1998)). Ces modèles sont aussi utilisés dans des outils pour la classification métagénomique (PhymmBL (Brady

et Salzberg, 2009), PHYSCIMM (Kelley et Salzberg, 2010)). TACOA (Diaz *et al.*, 2009), conçu pour la classification des fragments génomiques environnementaux, utilise l'approche des plus proches voisins.

Différents types de données biologiques non séquentielles sont analysés par des approches d'apprentissage automatique. Les technologies des puces à ADN (DNA microarray) et de séquençage des ARNs (RNA-seq) génèrent des quantités énormes de données d'expressions de gènes. Les méthodes d'analyse de l'expression de gènes utilisent des approches d'apprentissage supervisé (réseaux de neurones, bayésiens, *SVM*, *Random forest*, etc.), de *clustering* (k-moyennes, algorithmes basés sur la densité, etc.) ainsi de sélection d'attributs (par exemple des méthodes de sélection basées sur les SVM, la corrélation et le test de χ^2) (Pirooznia *et al.*, 2008). D'autres domaines de la bioinformatique font appel aux techniques d'apprentissage automatique telles que la prédiction de la structure secondaire des ARNs et protéines, la prédiction des sites de liaisons des facteurs de transcriptions, l'analyse de données de la chromatine et la classification fonctionnelle des gènes et des protéines (Tarca *et al.*, 2007; Libbrecht et Noble, 2015).

1.5 La classification des séquences : motivation et problématique

Le nombre des séquences nucléotidiques générées et entreposées dans les bases de données (privées et publiques) a connu une croissance exponentielle (figure 1.9). Cela est dû à l'accessibilité et aux faibles coûts des nouvelles technologies de séquençage. Ces séquences nucléotidiques sont de différentes tailles et correspondent aux séquences génomiques partielles et/ou complètes des milliers d'espèces étudiées sous différentes conditions biologiques.

Le séquençage des éléments et des marqueurs génétiques aide à l'élucidation et à la compréhension des différents mécanismes et voies biologiques tels que la croissance, le développement et la réponse aux stress biotiques (infections virales et

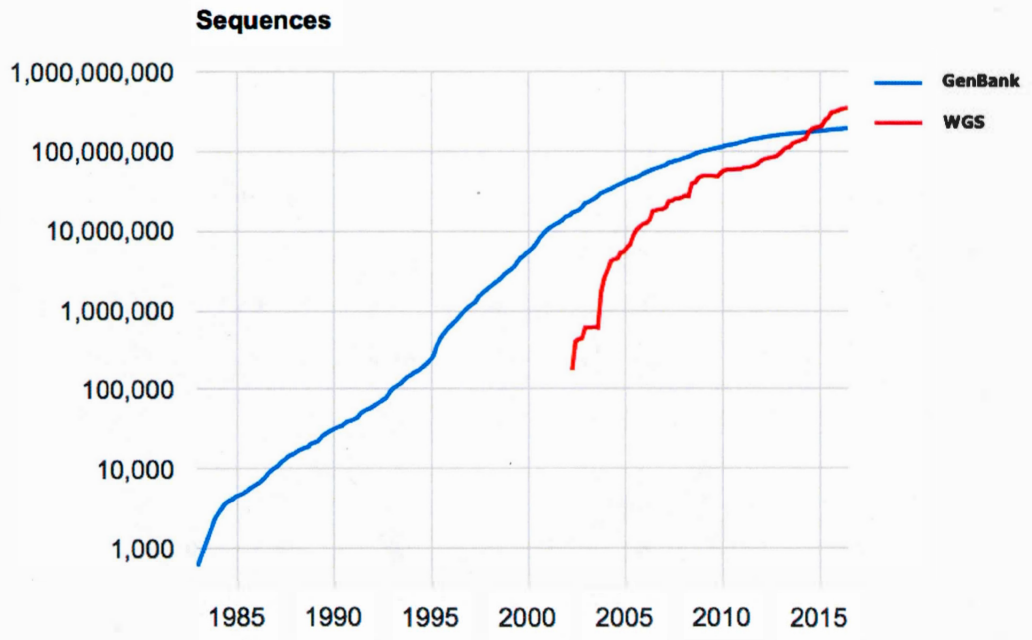


Figure 1.9: La croissance du nombre de séquences dans les bases de données GenBank et WGS de NCBI. Source : http ://www.ncbi.nlm.nih.gov/genbank/statistics

bactériennes) et abiotiques (froid, sécheresse, etc.). Aussi, la détermination de la composition génomique de chaque espèce est primordiale dans les études des processus évolutionnaires et des relations phylogénétiques. Par conséquent, les tâches d'identification, de classification et de mise en contexte génomique, biologique et taxonomique de chaque nouvelle séquence sont indispensables dans les différents domaines de recherche en biologie (animaux, plantes, microbiologie, etc.).

Nous nous intéressons dans ce mémoire à deux problématiques de classification distincts issues de deux projets majeurs et indépendants. Le premier projet consiste à la construction d'un système automatique d'annotation et de classification des

génomes viraux. Le deuxième projet cherche à identifier des miARNs exprimés par le blé et impliqués dans le développement et la réponse aux stress abiotiques (froid, salinité et aluminium) de la plante. Les deux projets ont généré un nombre énorme de séquences nucléotidiques ainsi des données de différents types apparentées à ces séquences et aux expériences. Les approches de classification abordées dans les deux projets sont différentes tenant compte de la divergence de leurs contextes et types de données. Cependant les deux approches sont conçues pour résoudre un problème standard de classification qui peut être reformulé comme suit :

Étant donné :

- Une séquence génomique S_i (partielle ou complète et nouvellement séquencée, extraite d'une séquence plus longue ou d'une base de données),
- Un ensemble de séquences de référence S étiquetées avec un ensemble de classes C,

Identifier :

- La classe C_α de la séquence S_i.

Cette formulation est conforme à un problème d'apprentissage supervisé. Les classes des séquences peuvent être des éléments génétiques, des fonctions biologiques, des caractéristiques, des traits, des rangs taxonomiques, etc. La classification des séquences constitue un problème difficile et non trivial.

En fonction des différentes classes ou organismes étudiés, plusieurs méthodes d'apprentissage automatique ont été développées. Elles peuvent être associées à une espèces (l'humain (Baumgartner *et al.*, 2004), la drosophile (Ohler *et al.*, 2002), le VIH (Struck *et al.*, 2014), etc.) ou génériques (Pirooznia *et al.*, 2008; Patil *et al.*, 2011). Selon la complexité des données, les approches de classification peuvent être simples (Edgar, 2010) ou intégratives combinant plusieurs méthodes dans un seul pipeline (de Oliveira *et al.*, 2005). Chaque étape du pipeline traite les

données de sortie des étapes précédentes afin de raffiner les résultats. Les méthodes de classification qui se basent sur la similarité des séquences utilisent un alignement de séquences (Huson *et al.*, 2007). Il existe d'autres méthodes, dites *ab initio*, qui ne dépendent pas de l'alignement de séquences. Elles utilisent des attributs calculés à partir des caractéristiques des séquences (par exemple la fréquences des motifs de taille fixe) pour faire entrainer un modèle d'apprentissage automatique (Wang *et al.*, 2007). Le choix des types d'attributs est important pour la construction d'un système discriminant. Généralement ces méthodes ont une efficacité proches à celles des méthodes basées sur l'alignement mais elles sont plus rapides et consomment moins de ressources.

Au delà du développement des méthodes d'apprentissage automatique, il est important en bioinformatique de fournir des services exploitant ces méthodes. Ainsi, plusieurs services dérivés des méthodes d'apprentissage automatique existent pour résoudre plusieurs problématiques tels que COMET pour la classification du VIH-1 (Struck *et al.*, 2014) et MiRdup pour l'identification des miARNs (Leclercq *et al.*, 2013). Afin de rendre ces méthodes réutilisables et reproductibles, il est important de permettre l'accès à ces services via des logiciels libres et interfaces web. Dans les deux projets abordés, l'une des exigences concerne l'accessibilité des services qui sont développés.

CHAPITRE II

A MACHINE LEARNING APPROACH FOR VIRAL GENOME CLASSIFICATION

Auteurs	Mohamed Amine Remita, Ahmed Halioui, Abou Abdallah Malick Diouara, Bruno Daigle, Golrokh Kiani and Abdoulaye Baniré Diallo
Journal	BMC Bioinformatics
État de l'article	Accepté le 15 mars 2017

2.1 Abstract

Background : Advances in cloning and sequencing technology are yielding a massive number of viral genomes. The classification and annotation of these genomes constitute important assets in the discovery of genomic variability, taxonomic characteristics and disease mechanisms. Existing classification methods are often designed for specific well-studied family of viruses. Thus, the viral comparative genomic studies could benefit from more generic, fast and accurate tools for classifying and typing newly sequenced strains of diverse virus families.

Results : Here, we introduce a virus classification platform, CASTOR, based on machine learning methods. CASTOR is inspired by a well-known technique in mo-

lecular biology : restriction fragment length polymorphism (RFLP). It simulates, *in silico*, the restriction digestion of genomic material by different enzymes into fragments. It uses two metrics to construct feature vectors for machine learning algorithms in the classification step. We benchmark CASTOR for the classification of distinct datasets of human papillomaviruses (HPV), hepatitis B viruses (HBV) and human immunodeficiency viruses type 1 (HIV-1). Results reveal true positive rates of 99%, 99% and 98% for HPV Alpha species, HBV genotyping and HIV-1 M subtyping, respectively. Furthermore, CASTOR shows a competitive performance compared to well-known HIV-1 specific classifiers (REGA and COMET) on whole genomes and *pol* fragments.

Conclusion : The performance of CASTOR, its genericity and robustness could permit to perform novel and accurate large scale virus studies. The CASTOR web platform provides an open access, collaborative and reproducible machine learning classifiers. CASTOR can be accessed at http://castor.bioinfo.uqam.ca.

keyword : sequence classification prediction virus classification

2.2 Background

Genomic sequence classification assigns a given sequence into its related group of known sequences with similar properties, traits or characteristics. It is a fundamental practice in different research areas of microbiology yielding major challenges in comparative genomics. Accurate genomic sequence classification and typing could help to enhance the phylogenetics and functional studies of viruses (Van Belkum *et al.*, 2001). They also help in determining pathogenicity, developing vaccines, studying epidemiology and drug resistance (Van Belkum *et al.*, 2001; Struck *et al.*, 2014). Recent advances in DNA sequencing and molecular biology techniques provide an immense collection of genomic information. Such data volume raises challenges for genetic-based classification techniques. Three main approaches have been designed and implemented to classify different types

of viruses based on their genomic sequence characteristics. The first is *sequence alignment-based* approach which is widely used, *e.g.* in similarity search methods (BLAST (Altschul *et al.*, 1997), USEARCH (Edgar, 2010), etc.) and in pairwise distance based-methods (PASC (Bao *et al.*, 2014), DEmARC (Lauber et Gorbalenya, 2012), etc.). The second is *phylogenetic-based* approach. It is implemented in several tools, *e.g.* REGA (de Oliveira *et al.*, 2005; Alcantara *et al.*, 2009) and Pplacer (Matsen *et al.*, 2010). The aim of these methods is to place an unknown sequence on an existing phylogenetic tree of a set of reference sequences. Each time a given sequence has to be classified, it is realigned with the set of reference sequences. Then, either a new phylogenetic tree is inferred or the given sequence is placed in the existing tree. The third is *alignment-free* approach including methods based on nucleotide correlations (Liu *et al.*, 2008) and sequence composition (Yu *et al.*, 2013; Struck *et al.*, 2014). It transforms sequences or their relationships to feature vectors and then constructs a phylogeny, a statistical model or a machine learning model (Vinga et Almeida, 2003; Bonham-Carter *et al.*, 2013). These methods are reviewed in Vinga et Almeida (2003), Mantaci *et al.* (2008), Xing *et al.* (2010) and Bonham-Carter *et al.* (2013). Restriction fragment length polymorphism (RFLP), a molecular biology technique (Williams, 1989), is used to type different virus strains (Bernard *et al.*, 1994; Nobre *et al.*, 2008; Janini *et al.*, 1996; Mizokami *et al.*, 1999; Nakao *et al.*, 1991). Several algorithmic approaches have tackled theoretical and experimental problems related to the restriction enzyme data such as restriction mapping problem (see Pevzner, 2000, chap. 2), phylogeny estimation (Adams et Rothman, 1982; Templeton, 1983; Felsenstein, 1992), SNP genotyping (Chang *et al.*, 2010) and analysis of RFLP digitized gel images (Bajla *et al.*, 2005; Maramis *et al.*, 2011). However, large scale computational sequence classification based on the RFLP technique is not yet covered in literature. Due to the genetic polymorphism in DNA sequences, fragments resulting from enzyme digestions are different in terms of number and length between individuals or

types. A set of restriction enzymes grounds a fragment pattern signature for each sequence. Therefore, similar sequences ought to have similar fragment patterns and thus similar restriction site distributions. This *a priori* knowledge could be used to build a machine learning model where sequences are represented by restriction site distributions as a feature vector and a class feature corresponding to a taxonomic level (genus, species, etc.). In this paper we introduce CASTOR, a machine learning web platform, to classify and type sequences. CASTOR integrates a new alignment-free method based on the RFLP principle. Our *in silico* method is independent of the sequence structure or function and is also not organism-specific. CASTOR is designed to facilitate the reuse, sharing and reproducibility of sequence classification experiments.

2.3 Methods

2.3.1 Overview of the approach

In this paper, we propose an *in silico* approach to identify and classify viral DNA sequences based on their restriction enzyme sites using supervised machine learning techniques. Like other supervised learning approaches, the proposed one is divided into two main units (Fig. 2.1). The *classifier construction unit* builds and trains classification models (or classifiers). It requires a set of reference viral genomic sequences, their classes and a list of restriction enzyme patterns. It starts by creating a training set including a group of feature vectors. The latter is computed from the distribution of the restriction site patterns on the given DNA sequences and then refined by feature selection methods. A collection of learning classifiers are then trained and evaluated using 10-fold cross-validation in order to choose the best classifier. The second unit (*prediction unit*) is intended to predict the classes or annotations of given viral sequences. The inputs of this unit are a classifier, a set of DNA sequences and the same list of restriction enzyme patterns used to

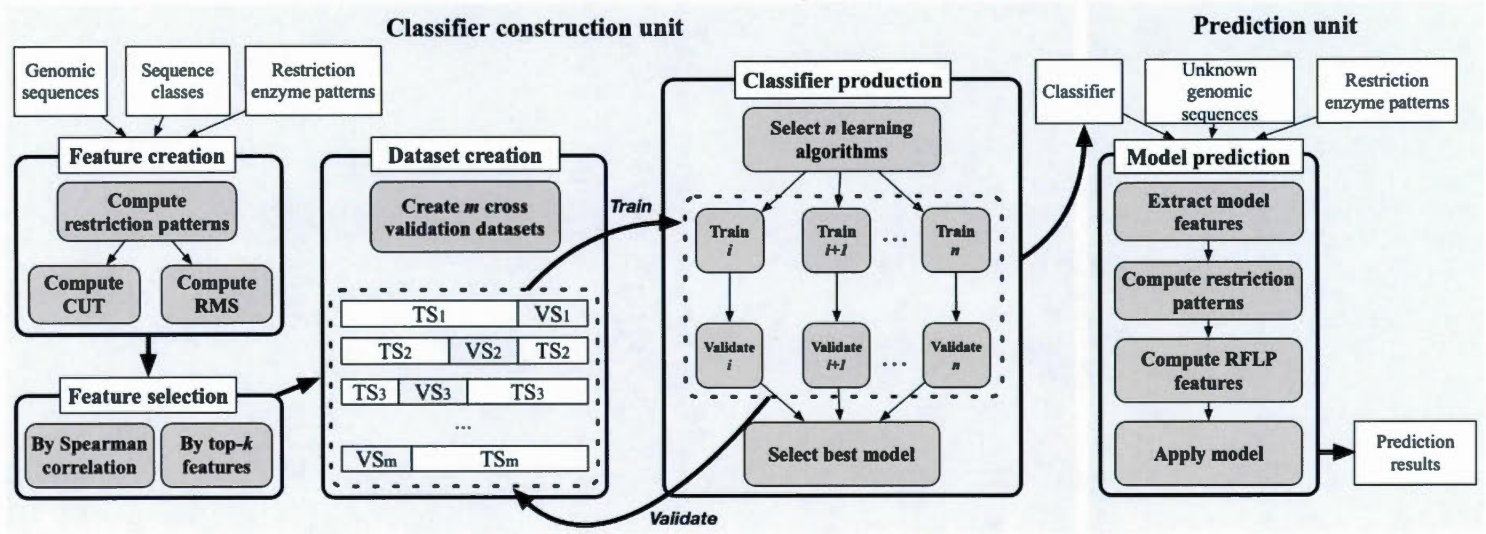


Figure 2.1: Overview of CASTOR kernel architecture. The kernel is composed of two main units (classifier construction and prediction). White rectangles represent input and output data; grey and curved rectangles represent processes. TS and VS are training set and validation set, respectively

train the classifier.

2.3.2 Restriction fragment pattern-based features

Here, we propose a set of features simulating the outcome of the RFLP technique. From REBASE database (Roberts *et al.*, 2015), we extracted a list of 172 type II restriction enzymes and their recognition sites. Type II family cleaves (cuts) DNA sequences precisely on each occurrence of the recognition site. Then, the restriction digestion of DNA sequences is computationally simulated. In order to build a training set, for a sequence s and enzyme z we compute two metrics representing the distribution of the digested fragments : the number of cuts of the enzyme ($CUT(s, z)$) and the root mean square of digested fragment lengths ($RMS(s, z)$) calculated as

$$RMS(s, z) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} l_i^2} \tag{2.1}$$

where n is the number of fragments ($CUT(s, z) + 1$) and l_i is the length of the i^{th} fragment in linear genomes. For circular genomes $n = CUT(s, z)$. Other metrics could be easily computed from the fragment digestion to construct the feature vectors.

2.3.3 Feature selection methods

The selection of an optimal subset of features improves the learning efficiency and increases the predictive performance. Feature selection techniques reduce the learning set dimension by pruning irrelevant and redundant features. Two relevant methods of feature reduction are provided. The first method (*topAttributes*) ranks the features according to their information gain (Ben-Bassat, 1982) and selects a subset of top-k features. Information gain estimates the mutual information between a feature and the target class. The second method (*correlation*) uses the

Spearman's rank correlation coefficient to construct a set of uncorrelated features. The correlation coefficient between two feature ranking vectors u and v of size n is computed as follows :

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (u_i - v_i)^2}{n(n^2 - 1)}. \tag{2.2}$$

A two-tailed *p-value* is computed to test the null hypothesis which states that two feature vectors are uncorrelated. In order to remove one of the two correlated features, two strategies could be used : discarding the feature with the largest sum of absolute correlation coefficients or the one with the smallest information gain score.

2.3.4 Learning and evaluation

We explored three types of classifiers : (1) symbolic methods (C4.5 decision tree (J48) (Quinlan, 1993) and random forests (RFT) (Breiman, 2001)), (2) statistical methods (naive Bayes classifier (NBA) (Langley *et al.*, 1992; John et Langley, 1995), support vector machine (SVM) (Cortes et Vapnik, 1995) and K-nearest neighbors (IBK) (Cover et Hart, 1967; Aha *et al.*, 1991)) and (3) ensemble methods (Adaboost (ADA)(Freund et Schapire, 1997) and Bagging (BAG) (Breiman, 1996) both combined with J48) ; see Table A.1 for more details. A 10-fold cross-validation strategy is used to assess the performance of the trained classifiers. Performance measures are weighted according to the number of instances and computed for the overall classification. The performance measures are :

$$TPR = TP/(TP + FN), \tag{2.3}$$

$$FPR = FP/(FP + TN), \tag{2.4}$$

$$Precision = TP/(TP + FP), \tag{2.5}$$

$$F - measure = \frac{2 \times TPR \times Precision}{TPR + Precision}. \tag{2.6}$$

where TP, TN, FP, and FN are the number of true positive, true negative, false positive and false negative predictions, respectively. TPR and FPR are the true positive rate and the false positive rate, respectively. We used Weka data mining program to perform the training and the evaluation (Hall *et al.*, 2009).

To include a negative class in the training sets, two approaches could be used. First, provide manually constructed negative class from collected relevant data. Second, build it with the provided negative class generator. This generator constructs altered sequences data from a sampling with replacement of the positive set sequences. To alter the sampled sequences, we reshape the RFLP length distribution of the training set by randomly shrinking, expanding or keeping unchanged the length of the sampled sequences. Then, each sequence is randomly shuffled while preserving k-mer counts.

2.3.5 Datasets

In this study, we applied our approach to a wide range of viruses. We selected one dsDNA virus (human papillomavirus (HPV)), one dsDNA-RT virus (hepatitis B virus (HBV)) and one ssRNA-RT virus (human immunodeficiency virus type 1 (HIV-1)). (1) HPVs have a circular double stranded DNA genome of ~8000bp and belong to five genera (Alpha, Beta, Gamma, Mu and Nu). HPVs belonging to a genus share over 53% identity of their complete genomes and ones in the same species level share over 62% of identity (Daigle *et al.*, 2015; Bernard *et al.*,

2010). We assessed the performance of HPV classification in the genus and species taxonomic levels. At the species level, we selected only the Alpha HPV genus representing the most abundant and diverse genomes in databases. It is divided into thirteen species (Alpha 1-11, Alpha 13-14). Unfortunately, some HPV genera (Mu and Nu) and Alpha HPV species (1, 5, 8, 11 and 13) were underrepresented and were therefore discarded. (2) HBV genomes are smaller (3200bp) and are circular partly double stranded DNA. HBVs are classified into eight genotypes (A-H) with at least 8% divergence among their genomic sequences (Schaefer, 2007). We evaluated the performances of our method for the genotyping of HBV strains. HPV and HBV complete genome sequences were downloaded from the NCBI RefSeq database (Coordinators, 2016). The taxonomic annotations were extracted from the NCBI Taxonomy database (Coordinators, 2016). (3) HIV-1 genomes have two copies of positive-sense single-stranded RNA with ~9700bp. Phylogenetically, HIV-1 strains are divided into four groups : M, N, O and P (Robertson *et al.*, 2000; Plantier *et al.*, 2009). M group strains are worldwide prevalent. They are categorized into pure subtypes (A-D, F-H, J and K) and recombinant forms (up to 70 CRFs and URFs). Genetic variations among subtypes are about 20-30% for *env* gene, 7-20% for *gag* gene and 10% for *pol* gene (Gao *et al.*, 1998). For HIV-1 classification, we studied complete genomes (CGs) and fragments covering *pol* gene from the position 2253 to 3554 with respect to HXB2 reference sequence and having a minimum size of 1Kbp (*pol* fragments). HIV-1 sequences were extracted from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/). For all the datasets, only complete, curated and well-annotated sequences were selected. Moreover, each class ought to have an adequate number of genomic sequences in order to have a representative genetic diversity.

2.3.6 Simulation studies

Raw viral sequence datasets, described above, were class-size imbalanced, *i.e.*, the difference in the number of genome sequences belonging to each class was relatively large. Generally, epidemiological studies are conducted on host-specific viruses (human, cattle, etc.) with the highest prevalence and pathogenicity (Muñoz *et al.*, 2003; Perz *et al.*, 2006). This leads to more data for some groups of viruses over others. Usually, training standard classifiers on imbalanced datasets affects their performance (mainly sensitivity and specificity) and misleads the interpretation of their accuracy (Libbrecht et Noble, 2015; Lin et Chen, 2013). Under-sampling majority class approach has been shown to perform well (Blagus et Lusa, 2010) and could be used with standard algorithms. Hence, from each previous dataset, we randomly performed under-sampling, without replacement, of the larger classes to have relatively the same sizes as the other classes. In order to identify the best parameters of the classifiers, we randomly sampled 10 datasets for each of the HPV genera, HPV Alpha species, HBV genotypes, HIV-1 M subtypes CGs and HIV-1 M subtypes *pol* fragments data. For each obtained sample, we performed a 10-fold cross-validation study with different classifiers built as follows. We constructed all the combinations of the two metrics (CUT and RMS), the two feature selection methods (*topAttributes* and *correlation*) and the seven learning algorithms. This construction yielded 28 $combinations * 10$ $datasets = 280$ $experiments$ for each virus classification.

2.4 Results and discussion

The results section is divided into four parts : first, we show how the RFLP signatures are suitable for viral classification ; second, we assess the performance of several competing classification algorithms on different virus datasets ; third, we compare the prediction made by CASTOR against widely used methods for

HIV-1 datasets, one of the most difficult to classify and fourth, we present the CASTOR web platform.

2.4.1 Classification with RFLP signatures in virus families

Figure 2.2 highlights the natural RFLP cuts in the collected HPV, HBV and HIV-1 datasets. The second column of the figure shows the multidimensional scaling (MDS) plot of the first two dimensions of distances between the feature vectors of the genomes. The separation between the different HPV genera (Fig. 2.2a) could approximatively be drawn, which is partly the case for the HPV species. The *Cohesion* (Daigle *et al.*, 2015) and *Silhouette* (Rousseeuw, 1987) indexes allow to measure the compactness and separability of classes. Here, both indexes show moderate values (between 0.2 and 0.8 for *Cohesion index* and -0.2 to 0.7 for *Silhouette index*) indicating that the classes are not well distinct. Several instances could be mislabeled or share the same RFLP cut patterns with other classes. This results in low or negative values of *Silhouette index* in HPV Alpha 3, 7 and HPV Gamma (Fig. 2.2a). With CASTOR, the best HPV Alpha Species classification obtains a TPR of 0.992 and FPR of 0.002 in 10-fold cross-validation analyses of 118 instances (see Table 2.1). The power of RFLP cuts in classification of viruses could be observed in HBV genotypes heatmap (see Fig. 2.2b). HBV highlights three genotypes (A, E and F) with *Cohesion indexes* for most instances above 0.7 indicating very coherent classes. But B and C genotypes have values between 0.1 and 0.6. The *Silhouette index* plots show several instances of B, C, E and G genotypes that have an striking disagreement with their assigned classes (*Silhouette index* < -0.1). Even with these constraints, CASTOR achieves the genotyping of 230 HBV instances with TPR of 0.996 and FPR of 0.001 according to a 10-fold cross-validation study (see Table 2.1). The HIV-1 cut site patterns have more variability among pure subtypes and CRFs (Fig. 2.2c). Likely, the MDS plot shows a moderate subtype clustering for the main HIV-1 subtypes. But

this clustering is not well separated compared to HPV and HBV. This variability among classes is reflected in low values of the *Cohesion index* (≤ 0.4). All, suggesting either variability, noise or mislabelling. For instance, $> 30\%$ of HIV-1 B and HIV-1 C instances tend to have RFLP cut patterns of other subtypes (negative *Silhouette indexes*). With CASTOR, the subtyping of HIV-1 group M within 18 main subtypes was assessed for 597 instances with a TPR of 0.983 and FPR of 0.001.

Previously, it has been clearly shown that RFLP has a power for classification in several viruses such as HPV (Bernard *et al.*, 1994; Nobre *et al.*, 2008), HBV (Mizokami *et al.*, 1999) and HIV (Janini *et al.*, 1996). But these studies are mostly limited to two to five classes. To the best of our knowledge, our study constitutes the first large scale and multi-class analyses of RFLP cut for classification. It provides the basis to explore large and various types of classifications, in particular those based on machine learning methods.

2.4.2 Machine learning classifier tuning and performance

The CASTOR platform relies on machine learning methods for the classification of viruses based on RFLP signatures of nucleotide sequences. The platform is detailed in the CASTOR web platform section. Three important parameters constitute the kernel of each CASTOR classifier : a metric, a feature selection method and a learning algorithm. To assess the different combination of the models, we performed a 10-fold cross-validation of the 280 experiments associated to each of the five datasets (HPV genera, HPV Alpha species, HBV genotypes, HIV-1 M subtypes CGs and HIV-1 M subtypes *pol* fragments). From the overall results of the five virus classifications, it is not obvious to distinguish the best candidate between CUT and RMS metrics. In the genotyping of HBV, CUT performs better than RMS ($p\text{-}value = 0.0012$, Wilcoxon/Kruskal-Wallis test) while in the HPV genera

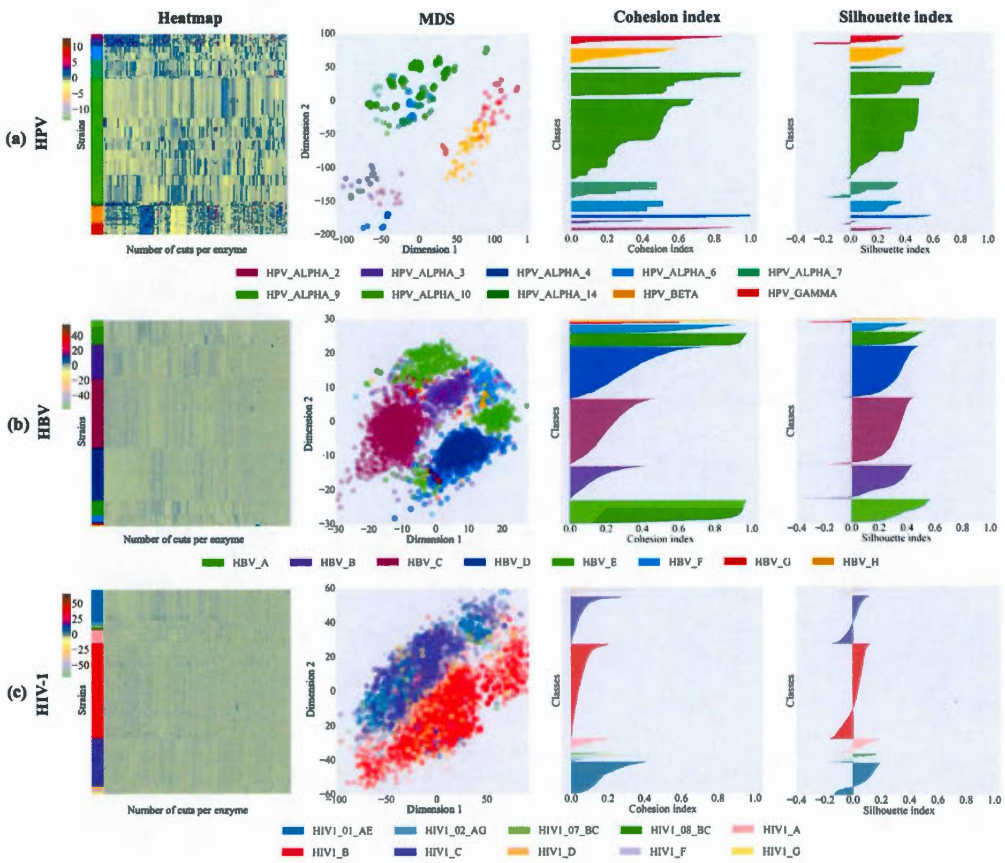


Figure 2.2: Class cohesion of three virus datasets. The four columns illustrate the separability and compactness of three virus complete genomes datasets based on 172 restriction enzyme cuts. The first column shows heatmaps of *CUT* clustered by x-axis. The samples in the y-axis are grouped by studied classes followed by intra-class clusterings. The second column shows MDS of the *CUT* distances between samples. The third and fourth column represent, respectively, the *Cohesion* and *Silhouette* indexes of the classes. (a) Classes in HPV are Alpha species, Beta and Gamma genera. (b) Classes in HBV are A-H genotypes (c) Classes in HIV-1 are M pure subtypes and CRFs

Table 2.1: CASTOR best accuracies on the classification of five datasets

Group of virus	Organism	Classification	# of classes	# of instances	*TPR*	*FPR*	*F-measure*	Classifier ID
I (dsDNA)	HPV	Genera	3	125	0.992	0.005	0.992	PMSHPV01
		Alpha species	8	118	0.992	0.002	0.992	PMSHPV02
VII (dsDNA-RT)	HBV	Genotypes	8	230	0.996	0.001	0.996	PMSHBV01
VI (ssRNA-RT)	HIV-1	Groups	4	76	1.000	0.000	1.000	PMSHIV01
		M Subtypes	18	597	0.983	0.001	0.983	PMSHIV02

This table contains the best results of the experimental study performed on the different datasets. The evaluation measures are obtained with 10-fold cross validation analysis. The column Classifier ID contains the corresponding models available in CASTOR platform.

and species classifications RMS performs better than CUT (*p-values* 5.00E-03 and 0.0293, respectively; Wilcoxon/Kruskal-Wallis test) (Fig. S1). However the mean of weighted *F-measures* for both methods are in all cases ≥ 0.906 (with a minimum of 0.793 and a maximum of 0.996). The same analyses were performed on HIV-1 CGs and *pol* fragments. CUT performs slightly better than RMS in both datasets when comparing the mean of weighted *F-measures* (*p-values* 0.0213 and 0.0237 for CGs and *pol* fragments, respectively; Wilcoxon/Kruskal-Wallis test). Due to the variability of HIV-1, the mean of weighted *F-measures* falls to 0.857 in CGs and 0.793 in *pol* fragments (Fig. S1). Hence for the remaining of our study, we will fix the RFLP metric according to its performance on the corresponding datasets.

Figure S2 presents the comparative analyses of the two feature selection methods (*correlation* and *topAttribute*) in the 280 experiments for each dataset. The mean of weighted *F-measures* of the two feature selection methods are not statistically different in all datasets (based on the Wilcoxon/Kruskal-Wallis test). In fact, the results of the two methods are correlated for the three viruses with the Spearman's rank correlation coefficient ranging between 0.772 and 0.968 (see Fig. S4). In these simulations, the seven learning algorithms have various performances according to the different datasets. The algorithm J48 has the worst weighted *F-measure* values (see Fig. 2.3). However, its performance improves when combined with RFT or BAG algorithms. In general, SVM performs better in four of five datasets with mean of weighted *F-measures* > 0.906 and ranks number one in HPV Alpha species, HBV genotypes and HIV-1 subtypes classifications and four in HPV genera classification. It is followed by RFT, NBA and IBK. However, RFT and NBA are affected by a large variance (Fig. 2.3). These rankings are clearly observable on Figure S3 and Figure S4 presenting respectively the correlations CUT/RMS and *topAttribute*/*correlation* grouped by algorithms. While most algorithms have si-

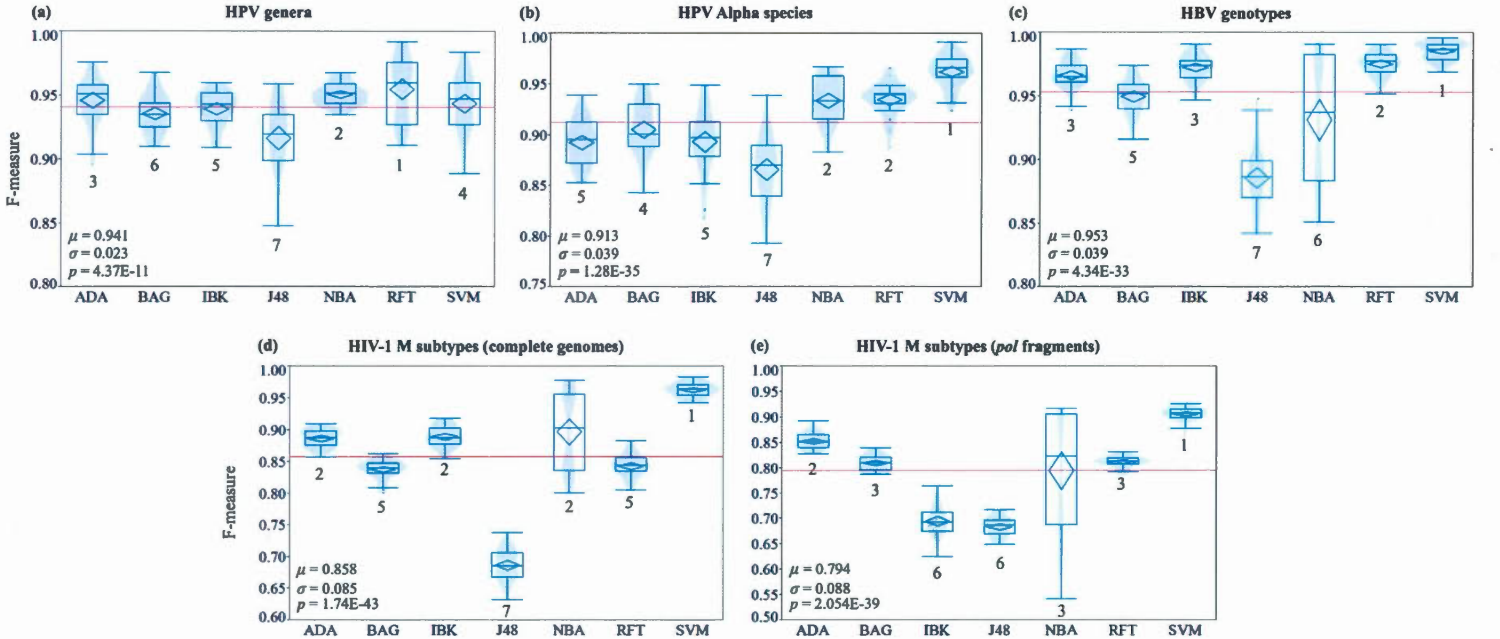


Figure 2.3: Learning algorithm evaluation on five datasets. This figure illustrates the *F-measure* distribution (boxplot) of seven learning algorithms on the prediction of (a) HPV genera, (b) HPV Alpha species, (c) HBV genotypes, (d) HIV-1 M subtypes with complete genomes (e) HIV-1 M subtypes with *pol* fragments. HPV and HBV datasets are complete genomes. The number below each boxplot corresponds to the statistically discriminative rank of the algorithms. The ranking is performed with paired Student's t test. μ, σ are the mean and the standard deviation of the overall *F-measures*, respectively. p is the *p-value* of the statistically significance of the weighted *F-measure* mean differences among the algorithms computed with the Wilcoxon/Kruskal-Wallis test

milar performance with *CUT* or *RMS*, Naive Bayes surprisingly performs better with *CUT*.

2.4.3 Assessing the performance CASTOR on HIV-1 data

2.4.3.1 CASTOR exhibits high accuracy for different HIV-1 classification

Table 2.2 highlights CASTOR prediction accuracies on five CG and seven *pol* fragment HIV-1 classifications. For each dataset, the best performing models (classifiers) have been identified according to a 10-fold cross-validation analysis. The *F-measure* of the best classifier for the HIV-1 groups M, N, O and P indicates that all the sequences are correctly classified (for CGs and *pol* fragments). For the prediction of the main HIV-1 pure subtypes as well as CRFs, *F-measures* are above 0.971 (with $FPR \leq 0.003$) for both CGs and *pol* fragments when the pure subtypes and CRFs are separate models. When combining pure subtypes and CRFs, the *F-measure* still remains above 0.971 for CGs but it drops to 0.919 when the classes are balanced to 30 instances per class or 0.962 for 200 instances per class. It appears that the CASTOR models are underperforming when we try to predict between pure subtypes and CRFs (*F-measures* of 0.795 and 0.885 for CGs and *pol* fragments, respectively).

2.4.3.2 Comparing COMET, REGA and CASTOR

Next, we compared the performance of CASTOR against the most powerful and widely used HIV-1 specific predictors namely COMET (Struck *et al.*, 2014) and REGA version 2.0 (de Oliveira *et al.*, 2005; Alcantara *et al.*, 2009) (Figure 2.4). These comparisons are based on CG as well as *pol* fragment data. It is important to notice that these programs are fixed and do not allow neither any change on the trained classes nor new training samples. Here the actual training of COMET and REGA includes respectively 55 and 22 classes for either CG or *pol* fragments.

Table 2.2: Evaluation of HIV-1 classification with CASTOR

	Classification	# of classes	# of instances	[min - max] instances/class	*TPR*	*FPR*	*F-measure*	Classifier ID
CGs	Groups (M, N, O and P)	4	76	[4 - 32]	1.000	0.000	1.000	PMVHIVGC01
	Pure subtypes	6	189	[30 - 36]	0.995	0.001	0.995	PMVHIVGC02
	CRFs	12	234	[10 - 30]	1.000	0.000	1.000	PMVHIVGC03
	Pure subtypes and CRFs	18	423	[10 - 36]	0.981	0.001	0.981	PMVHIVGC04
	Pure subtypes vs CRFs	2	200	[100 - 100]	0.795	0.205	0.795	PMVHIVGC05
pol	Groups (M, N, O and P)	4	94	[4 - 45]	1.000	0.000	1.000	PMVHIVPL01
	Pure subtypes	6	1800	[300 - 300]	0.983	0.003	0.983	PMVHIVPL02
	CRFs	16	480	[30 - 30]	0.971	0.002	0.971	PMVHIVPL03
	CRFs	6	1200	[200 - 200]	0.993	0.001	0.993	PMVHIVPL04
	Pure subtypes and CRFs	23	690	[30 - 30]	0.920	0.004	0.919	PMVHIVPL05
	Pure subtypes and CRFs	12	2400	[200 - 200]	0.962	0.003	0.962	PMVHIVPL06
	Pure subtypes vs CRFs	2	200	[100 - 100]	0.885	0.115	0.885	PMVHIVPL07

This table contains the *TPR*, *FPR* and *F-measure* of 12 HIV-1 classifications obtained with 10-fold cross validation analysis. For each classification, the number of corresponding classes and instances are given. The range [min-max] indicates the interval of instance frequencies per class used during the training of each model. The column Classifier ID contains the corresponding models available in CASTOR platform.

To avoid under-represented classes, CASTOR was trained on 18 classes for CGs and 28 classes for *pol* fragments (models are available under the classifier IDs PMSHIV02 and PMSHIV03, respectively). We performed three comparisons (see Figure 2.4). The first, named *complete sampling*, assesses the performance of each method on 10 percent of randomly sampled Los Alamos HIV data. This sampling permits to assess the performance of the predictors to fit realistic data with unknown classes. The second, named *specific subtypes*, focuses, for each method, only on the corresponding trained subtypes. The third, named *common subtypes*, compares the performance of the methods on the intersection of the 3 trained subtypes. This strategy is used due to the fact that the training of COMET and REGA cannot be changed. Thus, it is difficult to adapt or perform other classification studies or larger benchmark analyses. Figure 4 shows that for CGs, REGA performs the best followed by CASTOR and for *pol* fragments COMET outperforms, followed again by CASTOR. In the two types of data, when not performing the best, REGA or COMET performance drops drastically by more than 10% and ranks at the third position (Figure 2.4). Meanwhile CASTOR ranks second in both two types of data.

With CGs, CASTOR obtains a correct classification of 72.41% against the sampling of Los Alamos HIV data when REGA obtains 76.77%. But when testing predictors on their trained classes, the percentage of correct classification drastically increases to 98.33% and 96.61% respectively for REGA and CASTOR. This result remains almost the same when comparing only the common trained classes among the three predictors (Figure 2.4). These common classes cover 75% and 93% of the overall instances of the sampling of CGs and *pol* fragments, respectively. The mean TPR of CASTOR is higher than 0.950 in the case of either pure subtypes or CRFs. The TPR of REGA drops to 0.835 when assessing CRFs and remains almost perfect for pure subtypes (Table 2.3). In *pol* fragments, CO-

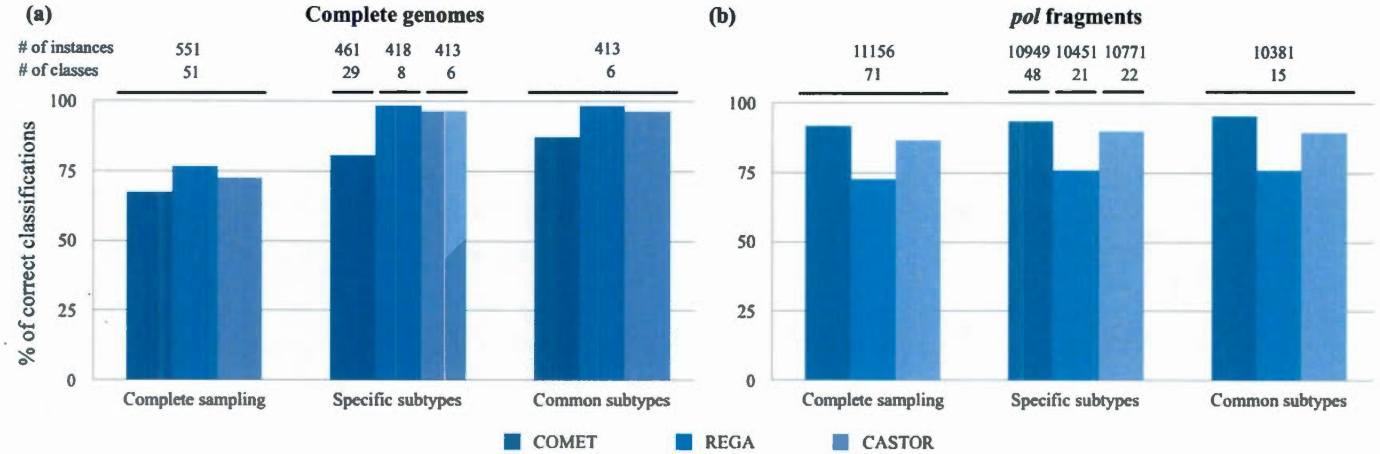


Figure 2.4: Performance of CASTOR with COMET and REGA predictors on HIV-1 datasets. The panels (a) and (b) show the percentage of correct classifications for HIV-1 complete genomes and HIV-1 *pol* fragments, respectively. The number of instances and the associated classes for each sampling is presented above the panels. Complete sampling corresponds to 10% of Los Alamos HIV data selected randomly. In specific subtypes sampling, the predictors are assessed against their trained classes. In common subtypes sampling, the predictors are assessed against the intersection of the classes of the three trained predictors

MET outperforms CASTOR and REGA in all comparisons. Applying the three methods, COMET, REGA and CASTOR, on 10% random sampling of Los Alamos HIV data, the percentages of correct classification were 91.74%, 72.48% and 86.64%, respectively. This result is confirmed when comparing only the common trained classes where COMET reaches 95.57% and CASTOR 89.51%. Note that REGA could not perform higher than 76% and has a mean TPR of 0.953 for pure subtypes competing with COMET. In CRF instances, COMET and CASTOR obtain almost an equal mean of TPR around 0.930 (Table 2.4). REGA cannot perform well in CRF classification and has a mean of TPR equal to 0.570. CASTOR has higher FPR values compared to the two other programs in overall classifications. This fact is not surprising since REGA and COMET are specifically tuned to predict HIV data. Their predictions with lower scores tend to be discarded or ambiguous. For instance, COMET has 32% of its CG predictions that are unassigned as well as 5% of its *pol* fragment predictions. Hence, these numbers are higher than the false positive values of CASTOR, but they are not included in the FPR computation. However, it will be interesting to include in CASTOR a threshold of inclusion of a given sequence into a class. This could help reducing the FPR but it would require deeper analyses. It also should be associated to the *open-set* classification problem that is beyond the scope of this paper.

Even though CASTOR is not a specific HIV-1 classifier, it competes with the most powerful methods in HIV-1. Unlike COMET and REGA, CASTOR provides an easy way of performing several types of classification (see Table 2.2). It also has no restriction on the size of data and is time efficient. Hence, we completed the analysis by performing a test on the whole Los Alamos HIV dataset (without the training sequences of the three methods). For CGs (3 778 instances), CASTOR completes the test in 1min59s with an accuracy of 91.2%. While for the *pol* fragments (119 005 instances), it requires 20min10s with an accuracy of 85.41%. It

Table 2.3: Performances of HIV-1 predictors on complete genome classification

		# of instances	COMET			REGA			CASTOR		
			TPR	*FPR*	*F-measure*	*TPR*	*FPR*	*F-measure*	*TPR*	*FPR*	*F-measure*
CRFs	HIV1_01_AE	100	0.960	0.000	0.980	0.970	0.000	0.985	1.000	0.000	1.000
	HIV1_02_AG	10	0.900	0.000	0.947	0.700	0.000	0.824	0.900	0.007	0.818
	Mean		0.930	0.000	0.964	0.835	0.000	0.905	0.950	0.004	0.909
Pure subtypes	HIV1_A	100	0.660	0.000	0.795	0.990	0.000	0.995	0.940	0.000	0.969
	HIV1_B	100	0.910	0.000	0.953	1.000	0.000	1.000	0.960	0.003	0.975
	HIV1_C	100	0.970	0.000	0.985	1.000	0.000	1.000	0.970	0.003	0.980
	Mean		0.847	0.000	0.911	0.997	0.000	0.998	0.957	0.002	0.975

This table contains *TPR*, *FPR* and *F-measure* of COMET, REGA and CASTOR on the prediction of HIV-1 M pure subtypes and CFRs complete genomes. The shown classes belong to the common subtypes sampling. The CASTOR model used in this evaluation is PMSHIV02.

Table 2.4: HIV-1 predictor performances on *pol* fragment classification

			COMET			REGA			CASTOR		
		# of instances	*TPR*	*FPR*	*F-measure*	*TPR*	*FPR*	*F-measure*	*TPR*	*FPR*	*F−measure*
CRFs	HIV1_01_AE	1000	0.989	0.000	0.993	0.007	0.000	0.014	0.956	0.001	0.975
	HIV1_02_AG	1000	0.952	0.002	0.967	0.000	0.000	0.000	0.853	0.005	0.897
	HIV1_06_cpx	698	0.924	0.000	0.958	0.938	0.000	0.965	0.927	0.003	0.943
	HIV1_07_BC	1000	0.977	0.000	0.988	0.988	0.000	0.993	0.982	0.002	0.980
	HIV1_08_BC	399	0.965	0.000	0.981	0.990	0.000	0.994	0.972	0.001	0.970
	HIV1_11_cpx	58	0.828	0.000	0.906	0.690	0.000	0.816	0.897	0.006	0.588
	HIV1_12_BF	222	0.860	0.000	0.925	0.374	0.000	0.544	0.932	0.008	0.807
	Mean		0.928	0.000	0.960	0.570	0.000	0.618	0.931	0.004	0.880
Pure subtypes	HIV1_A	1000	0.966	0.001	0.980	0.968	0.106	0.654	0.891	0.006	0.917
	HIV1_B	1000	0.995	0.001	0.993	0.945	0.000	0.970	0.817	0.007	0.866
	HIV1_C	1000	0.990	0.001	0.991	0.997	0.000	0.997	0.912	0.003	0.942
	HIV1_D	1000	0.938	0.000	0.968	0.911	0.000	0.953	0.892	0.010	0.899
	HIV1_F	1000	0.927	0.000	0.962	0.970	0.000	0.985	0.914	0.003	0.940
	HIV1_G	1000	0.915	0.001	0.952	0.929	0.007	0.931	0.778	0.003	0.860
	Mean		0.955	0.001	0.974	0.953	0.019	0.915	0.867	0.005	0.904

This table contains *TPR*, *FPR* and *F-measure* of COMET, REGA and CASTOR on the prediction of HIV-1 M pure subtypes and CFRs *pol* fragments. The shown classes belong to the common subtypes sampling. The CASTOR model used in this evaluation is PMSHIV03.

shows that CASTOR takes 0.01s to process a sequence that is far more efficient than the time results indicated in (Struck *et al.*, 2014) for REGA (28s/sequence), but 10-fold less efficient than COMET (0.001s/sequence) (Struck *et al.*, 2014). Furthermore, due to size issues, it is not possible to perform such large analyses in actual version of COMET server. Overall, CASTOR highlights a good accuracy on the classification of the three studied viruses. However this accuracy is slightly lower than specific virus predictors as shown previously. But it exhibits more analysis capacity, permitting several and highly accurate set of classifications. As shown in table 2.2, this accuracy is higher than 90% for almost all studies except for comparing HIV-1 M pure subtypes vs CRFs. For less complex genomes such as HPV and HBV, the mean of weighted *F-measures* is higher than 0.912. CASTOR will allow to increase the class representatives, to add or remove classes and also to benchmark several types of classification. For viruses without existing predictors, it could accurately cover the needs as it is for HPV, instead of relying on the similarity sequence search such as BLAST (Altschul *et al.*, 1997) or USEARCH (Edgar, 2010). Sequence search is generally not recommended for subtyping since it will not allow the identification of novel forms, it cannot also aggregate common attributes of a class while predicting (Struck *et al.*, 2014; Edgar, 2010).

2.4.4 CASTOR web platform

CASTOR is available as a public web platform. It is composed of four main applications. (1) **CASTOR-build** allows users to create and train new classifiers from a set of labeled virus sequences. It contains default parameters and advanced options letting users to customize the classifier parameters. It can be used also to update the parameters or input sequences of an already built classifier. The constructed classifiers can be saved in an exportable file locally or published to the community via CASTOR-database described below. (2) **CASTOR-optimize** constructs improved classifiers. Unlike CASTOR-build that allows users to define

metrics, algorithms and feature selection techniques, it assesses all combinations of the classification parameters and provides the best fitting classifier according to the input data. (3) **CASTOR-predict** is the kernel application that allows users to annotate viral sequences according to a chosen classifier. Also, it serves as an evaluation module for classifiers with labeled test sets. The results are provided with enriched graphics and performance measures (4) **CASTOR-database** is a public database of classifiers which allows the community to share their expertise and models. It facilitates experiment reproducibility and model refinement. A characteristic viewer and a search engine of the published classifiers are also implemented. Hence, from the interface of CASTOR-database, users can download, reuse, update and comment the classifiers. To the best of our knowledge, CASTOR constitutes the first RFLP-based prediction platform for the classification of viral sequences.

2.5 Conclusion

In this paper, we have shown that RFLP has a great performance in large scale sequence classification. We also provide CASTOR, the first viral sequence classification platform based on RFLP. We claim that CASTOR can perform well for different types of viruses (Group I, Group VI and Group VII) with mean of weighted *F-measures* > 0.900 in most cases (see Table 2.1). In the future, we will attempt to increase the performance by modelling the boundaries of the classes and including an *open-set* approach to deal with instances from unknown classes. The CASTOR platform implements several metrics and classifiers, allowing generic and diverse analyses within the same environment. CASTOR allows the storage of models enabling reproducible experiments and open data access. Even though CASTOR is scaled for viruses, it can be used and extended easily for other types of organisms, including whole genome and partial sequences. In the future, more models will be included, in particular those specialized in less studied or-

ganisms and/or without dedicated tools. In addition, scientists could add their tuned models helping CASTOR to enhance the predictions. We will also optimize the platform to allow other types of classification such as functional, disease related and geographical classifications. Hence, CASTOR could quickly become a reference in comparative genomics focusing on various types of sequence classification.

Availability of data and material

The CASTOR web platform is available at http://castor.bioinfo.uqam.ca.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

MAR, AH and ABD conceived and designed the study. MAR developed and implemented CASTOR program and web platform. MAR, AAMD, GK and ABD collected the virus datasets and analyzed the HIV-1 results. MAR and AH performed the simulations. MAR, BD and ABD computed and analyzed the *Cohesion* and *Silhouette* indexes. MAR and ABD analyzed the overall results. MAR, AH, AAMD and ABD wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds de Recherche du Québec-Nature et Technologie (FRQNT) to ABD. MAR, AAMD and BD are FRQNT fellows. MAR is a NSERC fellow.

CHAPITRE III

AN INTEGRATIVE APPROACH TO IDENTIFY HEXAPLOID WHEAT MIRNAOME ASSOCIATED WITH DEVELOPMENT AND TOLERANCE TO ABIOTIC STRESS

Auteurs	Zahra Agharbaoui, Mickael Leclercq, Mohamed Amine Remita, Mohamed A Badawi, Etienne Lord, Mario Houde, Jean Danyluk, Abdoulaye Baniré Diallo and Fathey Sarhan
Journal	BMC Genomics
État de l'article	publié le 24 avril 2015

3.1 Abstract

Background : Wheat is a major staple crop with broad adaptability to a wide range of environmental conditions. This adaptability involves several stress and developmentally responsive genes, in which microRNAs (miRNAs) have emerged as important regulatory factors. However, the currently used approaches to identify miRNAs in this polyploid complex system focus on conserved and highly expressed miRNAs avoiding regularly those that are often lineage-specific, condition-specific, or appeared recently in evolution. In addition, many environmental and biological factors affecting miRNA expression were not yet considered, resulting still in an incomplete repertoire of wheat miRNAs.

Results : We developed a conservation-independent technique based on an integrative approach that combines machine learning, bioinformatic tools, biological insights of known miRNA expression profiles and universal criteria of plant miRNAs to identify miRNAs with more confidence. The developed pipeline can potentially identify novel wheat miRNAs that share features common to several species or that are species specific or clade specific. It allowed the discovery of 199 miRNA candidates associated with different abiotic stresses and development stages. We also highlight from the raw data 267 miRNAs conserved with 43 miRBase families. The predicted miRNAs are highly associated with abiotic stress responses, tolerance and development. GO enrichment analysis showed that they may play biological and physiological roles associated with cold, salt and aluminum (Al) through auxin signaling pathways, regulation of gene expression, ubiquitination, transport, carbohydrates, gibberellins, lipid, glutathione and *secondary metabolism,* photosynthesis, as well as floral transition and flowering.

Conclusion : This approach provides a broad repertoire of hexaploid wheat miRNAs associated with abiotic stress responses, tolerance and development. These valuable resources of expressed wheat miRNAs will help in elucidating the regulatory mechanisms involved in freezing and Al responses and tolerance mechanisms as well as for development and flowering. In the long term, it may help in breeding stress tolerant plants.

Keywords : Abiotic stress Development Deep sequencing MiRNA prediction Expressed sequenced tags *Triticum aestivum. L* Vernalization

3.2 Background

Abiotic stresses such as cold, drought, salt and aluminum (Al) limit plant growth and development, causing reduction in crop yield and important economic losses for farmers. To tolerate these stresses, plants have evolved a broad spectrum of metabolic, physiological and developmentally adaptations. These adaptive changes

are under the control of dynamic networks of genetic regulatory mechanisms that involve a large number of stress responsive genes. MicroRNAs (miRNAs), a major class of small non-coding RNAs, have emerged as key regulators of gene expression at the post-transcriptional level during plant growth and development (Carrington et Ambros, 2003; Jones-Rhoades *et al.*, 2006; Xing *et al.*, 2011). Several studies have shown that many miRNA families are involved in response to different abiotic stresses in many species (Hsieh *et al.*, 2009; Sun *et al.*, 2012; Tang *et al.*, 2012; Zeng *et al.*, 2012). A large number of plant miRNAs and their targets have been identified in the plant model *Arabidopsis thaliana* and many other species. Recent results have shown that plant miRNA genes are dispersed throughout the genome (Rogers et Chen, 2013) within protein coding genes (Rogers et Chen, 2013; Rajagopalan *et al.*, 2006), introns of protein coding and non-coding genes, and in intergenic regions (Sunkar *et al.*, 2005; Szarzynska *et al.*, 2009). Moreover, miRNAs may be produced from repetitive transposable elements (Piriyapongsa et Jordan, 2008; Lucas et Budak, 2012). To date, at the best of our knowledge, 2707 wheat miRNA candidates were identified by both bioinformatics and experimental approaches, using wheat expressed sequence tags (EST) database, the available genomic sequences of the hexaploid wheat genome, its individual chromosome arms and its ancestors (Tang *et al.*, 2012; Lucas et Budak, 2012; Wei *et al.*, 2009; Xin *et al.*, 2011; Kantar *et al.*, 2012; Ling *et al.*, 2013; Wang *et al.*, 2013; Han *et al.*, 2013; Kurtoglu *et al.*, 2013; Meng *et al.*, 2013; Pandey *et al.*, 2013; Li *et al.*, 2013b; Deng *et al.*, 2014; Li *et al.*, 2014; Sun *et al.*, 2014; Han *et al.*, 2014; Pandey *et al.*, 2014; Yao *et al.*, 2007). Among the wheat miRNA published sequences, 237 are registered in miRBase, a database of experimental miRNAs (Kozomara et Griffiths-Jones, 2011), and 170 are registered in PMRD, a database of plant miRNAs identified using an *in silico* approach (Zhang *et al.*, 2010). Although the wheat genome is completely sequenced, it is not yet possible to perform a thorough genome-wide study in the hexaploid wheat *T. aestivum* since the genome is not

completely assembled and annotated. This is caused by its large and complex genome containing a high percentage of DNA repeats (hexaploid genome AABBDD with approximately 1.7×10^{10} bp with at least 80% of DNA repeats) (Brenchley *et al.*, 2012). *In silico* approaches for the prediction of miRNAs include screening genomic or EST databases for orthologous sequences of known miRNAs and analyzing their pre-miRNA hairpin structures. Although these approaches were successful in identifying conserved miRNAs in plants that have their genomes fully sequenced and annotated (Sunkar *et al.*, 2005; Adai *et al.*, 2005; Zhang *et al.*, 2005), they eliminate the potential of searching for low abundance miRNAs that are often lineage-specific (Fahlgren *et al.*, 2010) or condition-specific (Breakfield *et al.*, 2012) or that appeared recently in evolution (young miRNAs). The challenge is bigger using polyploid species with partially sequenced and assembled genome such as the hexaploid wheat having a high content of repetitive DNA. To tackle this issue, one should develop conservation-independent techniques based on structure analyses and/or expression pattern of dicer cleavage products among pre-miRNAs (Friedländer *et al.*, 2008).

Most computational approaches labeled as miRNA predictors are actually pre-miRNA predictors, in the sense that they identify candidate genomic regions that may form pre-miRNAs but rarely take into account the availability of candidate mature miRNA evidence within the pre-miRNA. Several tools such as miRDeep (Friedländer *et al.*, 2008, 2012), miRanalyzer (Hackenberg *et al.*, 2009, 2011) and MiRdup (Leclercq *et al.*, 2013) were developed to predict miRNAs from raw reads data and shown to be accurate in most cases. Furthermore, many factors that affect miRNA expression including genotypes, tissues, age, development stage, growth condition (soil, hydroponic solution, temperature, humidity and photoperiod), stress treatment, are rarely considered in previous wheat miRNA identification studies. All wheat reported miRNAs were identified in libraries produced

from seedlings or plants grown under normal conditions (Wei *et al.*, 2009; Meng *et al.*, 2013; Li *et al.*, 2013b; Sun *et al.*, 2014; Zhang *et al.*, 2010), or tissue exposed to heat (Xin *et al.*, 2011) or seedling (Pandey *et al.*, 2014) and pollen mother cells from plants (Tang *et al.*, 2012) exposed to cold stress (Tang *et al.*, 2012), or drought (Kantar *et al.*, 2012). They were identified from different genotypes of winter or spring wheat in soil, or hydroponic solution and under different photoperiod conditions, or in field conditions. Since miRNA expression is tissue specific and regulated in response to plant development and growth conditions, the miRNA repertoire of hexaploid wheat is still incomplete. Although a large number of miRNAs associated with development or some abiotic stresses in wheat were previously identified, their functional diversity in Al, freezing tolerance, and floral transition in winter wheat is still unknown. Hence, the identification of miRNAs associated with tolerance to abiotic stress and floral transition is a first step towards the elucidation of their role in wheat.

To ensure an accurate identification of a large fraction of miRNAs associated with different physiological conditions in both stress sensitive and tolerant wheat, we conducted the present study to : 1) identify miRNAs from different tissues of plants from different genotypes grown under different stress conditions (cold, salt and aluminum) and at different development stages (vegetative and reproductive phases) ; 2) develop an integrative pipeline that combines bioinformatic tools, biological insights about known miRNA expression and dicer ligation patterns according to the universal plant miRNA criteria (Friedländer *et al.*, 2008, 2012), miRNA expression profiles in deep sequencing data (Meyers *et al.*, 2008), functional classification and experimental approaches (Figure 3.1). The bioinformatic tools include Mipred (Jiang *et al.*, 2007), HHMMiR (Kadri *et al.*, 2009), MIRcheck (Jones-Rhoades et Bartel, 2004) and MiRdup*, a plant updated version of our machine learning MiRdup (Leclercq *et al.*, 2013) which validates the position

of sequenced miRNAs in its corresponding folded pre-miRNA. Our integrative approach allows the discovery and profile of 165 novel hexaploid wheat abiotic stress responsive candidate miRNAs including ones associated with cold (52 miRNAs) and Al (27 miRNAs) tolerance as well as 99 developmentally responsive miRNAs with a high confidence level. It is the first study to report a large scale identification of hexaploid miRNome miRNAs from different tissues of sensitive and tolerant genotypes under normal conditions and short/long exposure to different abiotic stresses during vegetative and/or reproductive phase.

3.3 Results

3.3.1 Identification of miRNA candidates and their targets in hexaploid wheat

Our miRNA discovery pipeline consists of more than twenty steps divided in three main parts : producing and sequencing small RNAs, predicting miRNAs from deep sequencing data, classifying predicted miRNAs based on their expression profiles and Gene Ontology (GO) of their target genes (Figure 3.1). The sequencing of ten constructed libraries from three wheat genotypes grown under different abiotic stress conditions and development stages (Method B.1.1) yielded a total of 89,105,096 redundant raw reads (66,400,401 distinct reads) with 56% of high sequence quality (Table B.4). Before mapping the raw reads, we collected 1.4 million wheat ESTs from several databases, clustered into 127,039 Uniref clusters yielding to the best of our knowledge, the largest well-annotated EST databank in wheat (Method B.1.2). After raw reads adapter removal, a total of 56.4 million unique raw reads were mapped to our collected EST database resulting of 5.4 million unique mapped sequences (Figure 3.1). We identified 168,834 small RNAs and extracted 337,668 potential pre-miRNAs. Among the extracted pre-miRNA, 17,180 and 39,144 potential hairpins satisfy the minimal structural criteria of miRNAs assessed by two well-known pre-miRNA predictors, MiPred (Jiang *et al.*, 2007)

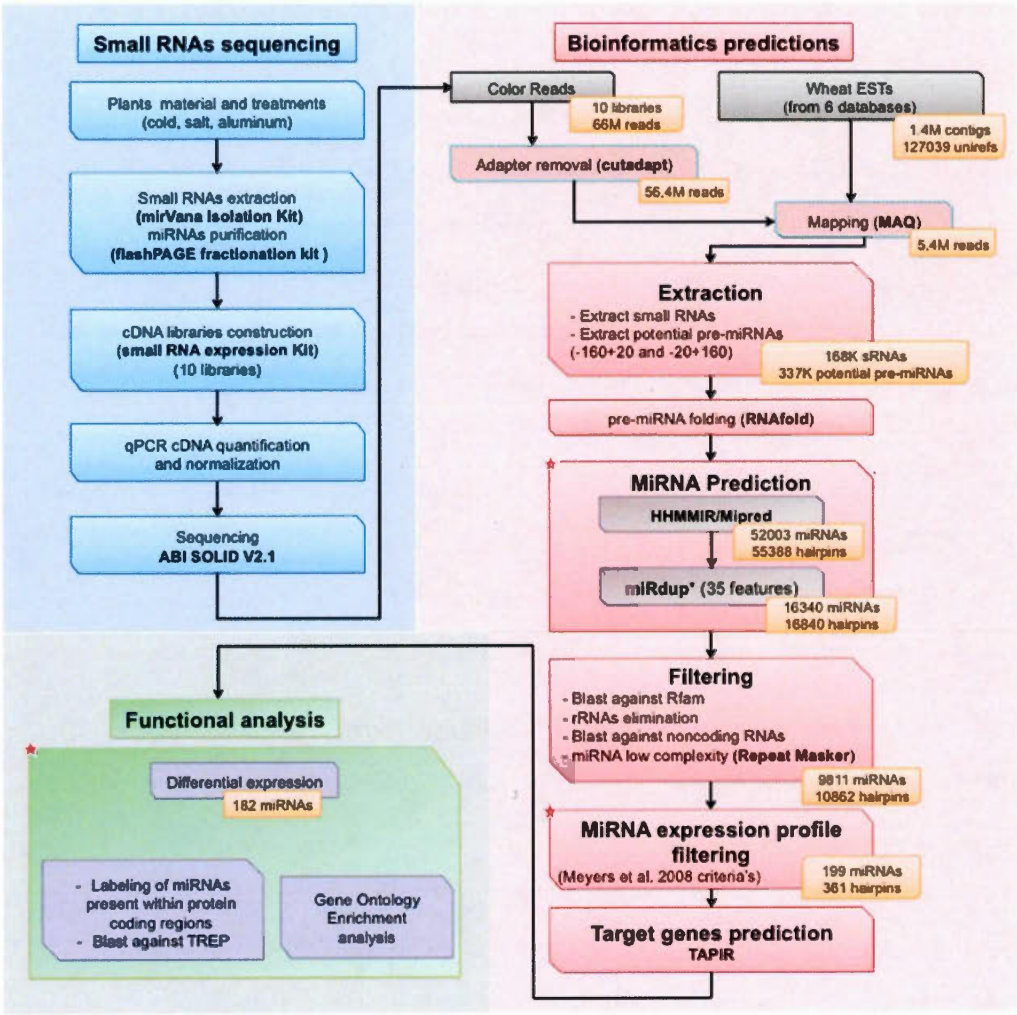


Figure 3.1: Overview of the wheat miRNA pipeline. The procedure is divided in three parts : producing and sequencing small RNA libraries, the bioinformatic prediction of miRNAs and functional analysis of the predicted miRNAs. The customized or developed steps are marked by stars. Orange boxes specify the data at hand after each given step. For details see Experimental procedure.

and HHMMiR (Kadri *et al.*, 2009), respectively (Figure 3.1). The lack of consideration of the mature miRNA localization in the hairpin, the expression profiles of reads throughout the hairpin, and/or other evidence of plant-miRNA characteristics represent major causes of overestimation of the number of candidate hairpins. We used a two-step method of miRNA identification within its hairpin from the folded pre-miRNA using MiRdup* and expression profile filtering using Meyers *et al.* relaxed criteria (Meyers *et al.*, 2008) (Figure 3.1). In the first step, MiRdup* trained on experimentally validated miRNAs from different datasets (all miRBase species, all plants and monocots only) localizes miRNA positions in pre-miRNAs (Method B.1.4 and Table B.5). The use of MiRdup* reduced the number of pre-miRNA hairpin candidates by 81% (Figures B.1 and B.2). In the second step, the expression profile filtering was based on miRNA expression pattern using the abundance of the miRNA candidates in each library and the distribution of reads mapped to a candidate pre-miRNA according to Meyers *et al.* relaxed criteria (Meyers *et al.*, 2008). This allowed the elimination of the miRNA dicer-Like candidates and the reduction of miRNA candidates by 84%.

Overall, this method results in the identification of more candidates (Figures B.1 and B.2) compared to Meyers *et al.* (2008) and MIRcheck (Jones-Rhoades et Bartel, 2004). Consequently, it yields pre-miRNA candidates that have various ranges of secondary structures as shown in dotbracket notation (Data B.2.1). Taken together, our approach identifies 199 unique miRNA candidates associated with 361 pre-miRNAs (Data B.2.1). It is important to notice that the majority of reads (95%) and the predicted miRNAs (64%) have the highest quality value for their sequence (Table B.6). It is important to note that, MiRdup* captures 95% of the miRNAs identified using MIRcheck with the same criteria from Meyers *et al.* (2008) while 151 putative candidates (containing validated candidates) are excluded by MIRcheck (Figures B.2). In addition, our pipeline identified miRNAs that

have features that are species specific, clade specific or shared between several species. We found that among the 199 identified miRNAs, 147 were identified by MiRdup* trained on all species of miRBase; only 49 were commonly identified by MiRdup* trained on the three datasets (all miRBase, all plants, monocot only). This suggests that these miRNAs share common features with all the widely separated plant lineages recorded in the database miRBase. For instance, apMir_22246 corresponding to miR160 with perfect match in wheat and moss *Physcomitrella patens* (ppt-miR160) is highly expressed in our investigated conditions indicating that this miRNA may play common biological functions in plants kingdom. While 109 miRNAs were only identified by MiRdup* trained on all plant and 92 miRNAs when trained on only monocot.

The number of identified miRNAs may be an overestimation due to the redundancy created by similar but not identical ESTs in part due to the polyploid nature of wheat. In the latter scenario, two or more closely related ESTs (true homeologs or ESTs with SNP differences) could encode identical or closely related miRNAs. Furthermore expressed isomiRNAs that share the same properties with the real miRNA in one library could be the dominant functional in another library. The Figure B.3a highlights that the majority (about 69%) of the predicted miRNAs are associated with one pre-miRNA. Furthermore, most of the pre-miRNA candidates (93%) harbour a unique miRNA leading to an exclusive miRNA/pre-miRNA association (Figure B.3b). To characterize further the nature of the pre-miRNA candidates, we determined if they were associated with repetitive transposable elements and protein coding regions. The results revealed that 20% of pre-miRNAs that correspond to 6.5% of miRNA candidates overlap with transposable elements (at e-value of 5E-5 with 80% identity) from TREP database (Table B.7a). In addition, 15% of ESTs corresponding to less than 5% of miRNA candidates overlap partly with protein-coding regions (at e-value of 1E-20 with

75% identity) from protein plant database (Table B.7b).

Prediction of miRNA targets is an important step to elucidate miRNA function in regulating gene expression. Among the identified candidates, 67% (133/199) were predicted to have Uniref annotated target genes (Table B.8). Unlike animal target genes, it is generally accepted that plant targets adopt a perfect seed match with the corresponding miRNAs, allowing more accuracy in their prediction. We found that 37 miRNA candidates are predicted to target a unique gene identified as UniRef (Figure B.3c). Although the majority of miRNA candidates seem to have more than two targets, detailed analysis reveals that in many cases the targets annotated with different UniRefs have the same gene description (Table B.9). To better explore the functional properties of the target genes, GO analyses were performed (Figure B.3d). We computed the enrichment of main GO Slim terms found within these targets based on the three GO categories (Figures B.4a-B.6a). Table 3.1 shows the enriched GO Slim terms and relevant associated target genes in libraries. An extensive description of GO enrichment analysis is presented in Appendix B.3. Our results revealed that miRNA candidates target regulators, cell metabolism and transport genes. The regulatory genes are enriched for many transcription factors and protein families (Table 3.1 and Table B.9). They are involved in regulation of gene expression, signal transduction pathways and ubiquitin-mediated protein modifications (GO Slim *Nucleus* with $P-value = 9.1e-004$, GO Slim *DNA binding* with P-value from $1.0e-003$ to $1.0e-005$, GO Slim *DNA metabolic process* with $P-value = 5.5e-004$, GO Slim *protein binding* with $P-value = 1.2e-006$). Cellular metabolism genes are involved in hormone, lipid and carbohydrates metabolism (GO Slim *catalytic activity* with $P - value = 3.4e - 006$), amino acid metabolism (GO Slim *secondary metabolic process* with $P - value = 3.7e - 008$) (Table B.9).

Table 3.1: Selected GO Slim enrichment in the different libraries and their relevant target genes

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	All	GO Slim description	Relevant associated targets
CC			++									Nucleus	CBFIVb-B20, MIKC-type MADS-box transcription factor, Auxin-responsive protein, Homeobox-leucine zipper protein, Histones, E3 ubiquitin-protein ligase
MF			+		++			+	++	++		DNA binding	DEAD-box ATP-dependent RNA helicase, Homeobox-leucine zipper proteins, Histones
					+++							Catalytic	Gibberllin 20 oxidase, Lipoxygenase, Glutathione peroxydase, superoxyde dismutase, Fructose-bisphosphate aldolase, Alcohol dehydrogenase, NADP-malic enzyme
				++			++					Transferase	Caffeic acid 3-O-methyltransferase, Tricetin 3',4',5'-O-trimethyltransferase, Serine/threonine-protein kinase, Glutathione-S-transferase

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	All	GO Slim description	Relevant associated targets
	+++	+++		+++		+++	+++				+++	Lipid binding	Non-specific lipid transfer protein, Homeobox-leucine zipper proteins
	++	++			++			+++	+++	++	+++	Protein binding	Histones, WCOR719, DNA-directed, RNA polymerase, MIKC-type MADS-box transcription factors, Serine/threonine-protein kinase, Cullin-1
BP	+	+										Transport	Sodium-hydrogen exchanger, Triose phosphate translocator, Protein transport protein, non-specific-lipid transfer protein, Sucrose transporters
	++		+++	++	+	+	++	++	++	++		Resp. to endogenous stimulus	Auxin-responsive protein, Homeobox-leucine zipper protein, Myo-inositol 1-phosphate synthase

L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	All	GO Slim description	Relevant associated targets
		++				+					Multicellular organismal dev.	Ubiquitin-like-specific protease, MIKC-type MADS-box transcription factors, Homeobox-leucine zipper protein
	++				+					++	Flower dev.	DL related protein, Ubiquitin-like-specific protease
								++			DNA metabolic process	Replication factor C, Histones, BARE-1 polyprotein
				+++						++	Sec. metabolic process	Phenylalanine ammonia-lyase, Tricetin 3',4',5' -O-trimethyltransferase

The enrichment is presented in four different symbols ("+++" for high (P-value $< 10^{-5}$), "++" for medium (P-value $< 10^{-3}$), "+" for low (P-value < 0.05) and *"nosymbol"* for no enrichment (P-value $= 0.05$). CC, cell component, MF, molecular function, BP, biological process. For details about the libraries and investigated conditions see Method B.1.1; and about GO Slim terms classification and associated target genes, see Figures B.4a-B.6a.

3.3.2 Characteristics of the miRNA candidates

Among the 199 predicted miRNAs, 30 have sequence homology with 76 published miRNA in wheat (Figure 3.2a). In addition, we explored the potential of conserved miRBase families present in our raw data that could not be mapped into the EST and found 267 miRNAs, corresponding to 43 families from which 25 families are known in wheat (Lucas et Budak, 2012; Wei *et al.*, 2009; Xin *et al.*, 2011; Kantar *et al.*, 2012; Wang *et al.*, 2013; Han *et al.*, 2013; Kurtoglu *et al.*, 2013; Meng *et al.*, 2013; Pandey *et al.*, 2013; Li *et al.*, 2013b; Zhang *et al.*, 2010; Dryanova *et al.*, 2008; Jin *et al.*, 2008; Yin et Shen, 2010) and 18 families have homology with many plant species (Figure 3.2b). It is important to notice that 13 families (58 miRNAs) have not yet been reported in previous wheat miRNA identification studies recorded in miRBase (Figure 3.2a). One can notice that the expression patterns of predicted miRNAs are different between the 10 libraries (Figure 3.2c). The abundance of 160 miRNAs corresponds at medium expression (abundance between 100–999 reads) in at least one library and 39 miRNAs have high reads abundance. MiRNAs were also classified according to their expression proportion over the total reads mapping to the corresponding pre-miRNAs (Jeong *et al.*, 2011). Hence, one class would correspond to *typical miRNA* when its expression represents more than 50% of the expressed small RNAs mapping a given pre-miRNA (Table B.10 and Data B.2.2) (Cuperus *et al.*, 2011). Above 80% of the predicted miRNAs in each library correspond to typical miRNAs (Table B.10) and they correspond to highly confident expressed miRNAs (Data B.2.2).

The diversity of predicted miRNA sequences is greater at 19 nt in length in all libraries (Figure B.7a-d) while the diversity of conserved miRNAs is greater at 21 nt (Figure B.7e) (unique sequences). For redundant miRNA sequences, a major peak at 21 nt was observed for both predicted and conserved miRNAs (Figure B.7a-e).

In addition, the majority of the identified miRNA expressed in the different investigated conditions are 19 or 21 nt long depending on the tissues, stresses, growth conditions or genotypes (Figure 3.2d). These miRNAs were shown to regulate at least 150 targets (60 unirefs) and at most 900 targets (335 unirefs) in all the explored conditions (Figure 3.2e).

3.3.3 Confirmation of predicted miRNA candidates

Selected miRNA candidates were validated by northern blotting, a useful criterion for authenticating miRNAs (Ambros *et al.*, 2003). For this selection, miRNAs were ranked according to their expression level. Then, candidates were randomly chosen from either the predicted only by MiRdup* (three tested cases) or predicted by both MiRdup* and MIRcheck (six tested cases) as well as low, medium and high expression. Their characteristics and their secondary structures are presented in Figure 3.3a and Table 3.2. These structures reveal the less stringent rules in MiRdup* concerning the symmetric and/or asymmetric bulges in which the number of successive unpaired bases could range up to five nt in the duplex such as in apMiR_16808 (Figure 3.3a). Their expression was confirmed under all the investigated conditions (Figure 3.3b and c). Many probes detect more than one mature miRNA product with distinct lengths in different libraries, 19/21 nt for apMir_14769, 21/23 nt for apMir_20602, and apMir_22246 (Figure 3.3b and c). This indicates that the second detected miRNA product may be a variant of each of these miRNA candidates. At least two of these miRNAs exhibit complex expression patterns in response to cold, vernalization, salt, Al, and in development (Figure 3.3b). For instance, the larger miRNA product detected for apMir_14769 is preferentially expressed in the Al-treated library from spring wheat (L8). In addition, in some libraries the expression level of the apMir_20602 and apMir_22246 is much higher than what may be expected from the low read numbers obtained from deep sequencing (Figure 3.3c and Data B.2.1). This may be due to the pre-

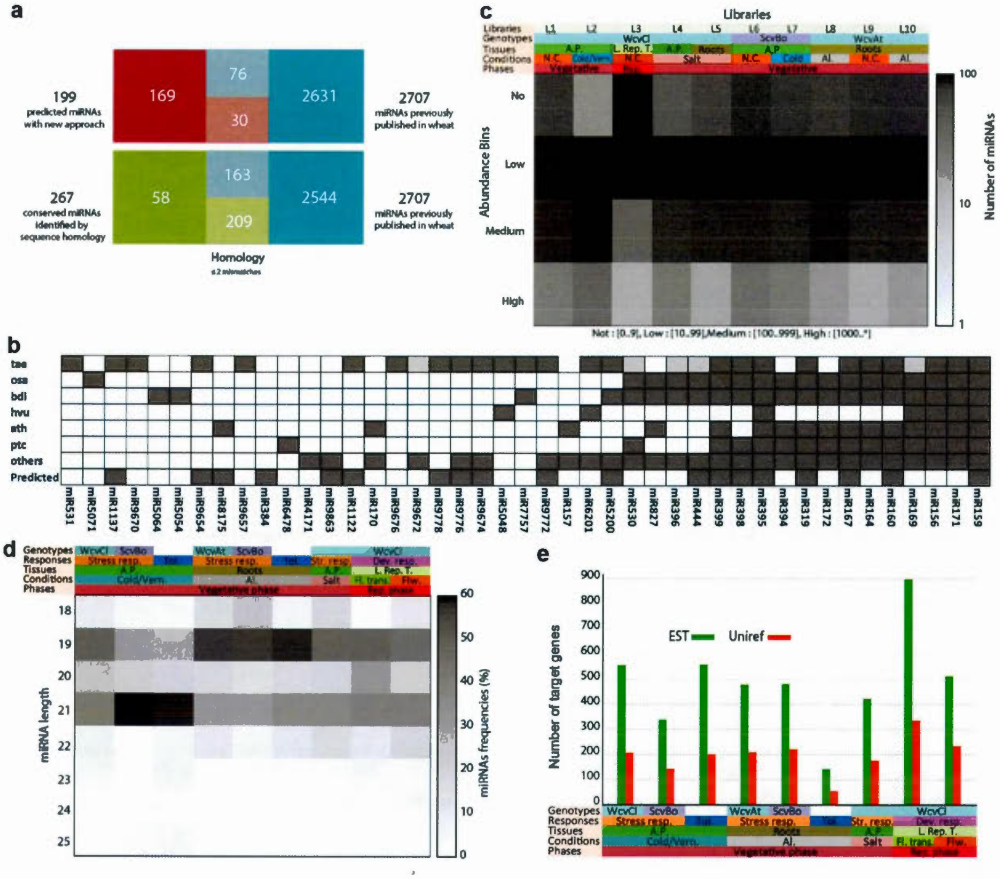


Figure 3.2: Overview of the predicted miRNAs. **a)** Diagram of the intersection between miRNAs predicted by the novel approach, conserved miRNAs identified by sequence homology, and miRNAs published in the literature; **b)** Evidence of conserved plant miRNA families present in miRBase including those predicted by our approach (tae, osa, bdi, hvu, ath and ptc correspond respectively to *Triticum aestivum, Oryza sativa, Brachypodium distachyon, Hordeum vulgare L., Arabidopsis thaliana and Populus trichocarpa*); **c)** The abundance bins of all predicted miRNAs in the 10 libraries (L1-L10) produced from plants grown under different investigated conditions. The abundance of the identified miRNAs represents the number of reads sequenced in each library and classified on 4 levels : low, 10–99 reads; medium, 100–999 reads; and high, 1000 and up; **d)** the length distribution of miRNAs associated with miRNAs differentially expressed in different investigated comparisons; **e)** the number of miRNA targeted genes (presented by EST id or UniRef id) associated with miRNAs differentially expressed in different investigated comparisons. WcvCl, winter wheat cultivar Clair (cold tolerant); ScvBo, spring wheat cultivar Bounty (cold and Al sensitive); WcvAt, winter wheat cultivar Atlas (Al tolerant); A.P, aerial parts; L./Rep. T., leaves and reproductive tissues; N.C, normal conditions; Al, Aluminum; Vern., vernalization; Rep., reproductive; Str. Resp., stress response; Tol., tolerance; Dev. resp., developmentally response; Fl. Trans, floral transition; Flw., flowering. See Method B.1.1 and Table B.3 for libraries and conditions and Table B.12 for the different investigated comparisons.

sence of very closely related miRNA variants that can hybridize with the probes especially if the mismatches are at their start/end. Probes used would not be able to differentiate between these possibilities and thus would represent an average response of these related miRNAs. The miRNA size may affect an AGO1 functional state that mediates the recruitment of RDR6 (Cuperus *et al.*, 2011; Chen *et al.*, 2010). However, for the apMir_20602 whose precursors overlap with transposable elements (Table B.7a), the high expression level and the presence of more than one size detected by northern may be associated with their repetitive nature with sequence variation in the genome.

3.3.4 Expression of the identified miRNAs in response to different abiotic stresses and plant development in wheat

To identify miRNAs associated with short and long exposure to cold, salt and Al responses and tolerance, three different control and five treated libraries from sensitive and tolerant wheat genotypes were used. To identify miRNAs associated with floral transition and flowering in winter wheat, one library from plants at vegetative phase under normal growth conditions, one library under vernalization conditions (long exposure to cold acclimation) and one library from de-acclimated (one week under favourable conditions after cold acclimation) plants at the reproductive phase were used. Analysis of miRNA expression levels identified 91% (182/199) of miRNAs that are differentially expressed between the stress conditions compared to the control by more than twofold change with a FDR of 0.05 (see an example of volcano plot showing differential expression of miRNA candidates in response to long exposure to cold in the Figure 3.4a). Out of the 182 miRNAs, 165 miRNAs are responsive to different abiotic stresses (cold, Al and salt) and 99 miRNAs are associated with plant development, particularly floral transition and flowering (Figure 3.4b and Table B.11). Among abiotic stress responsive ones, 52 and 27 miRNAs are associated with cold and Al tolerance, respectively (Table

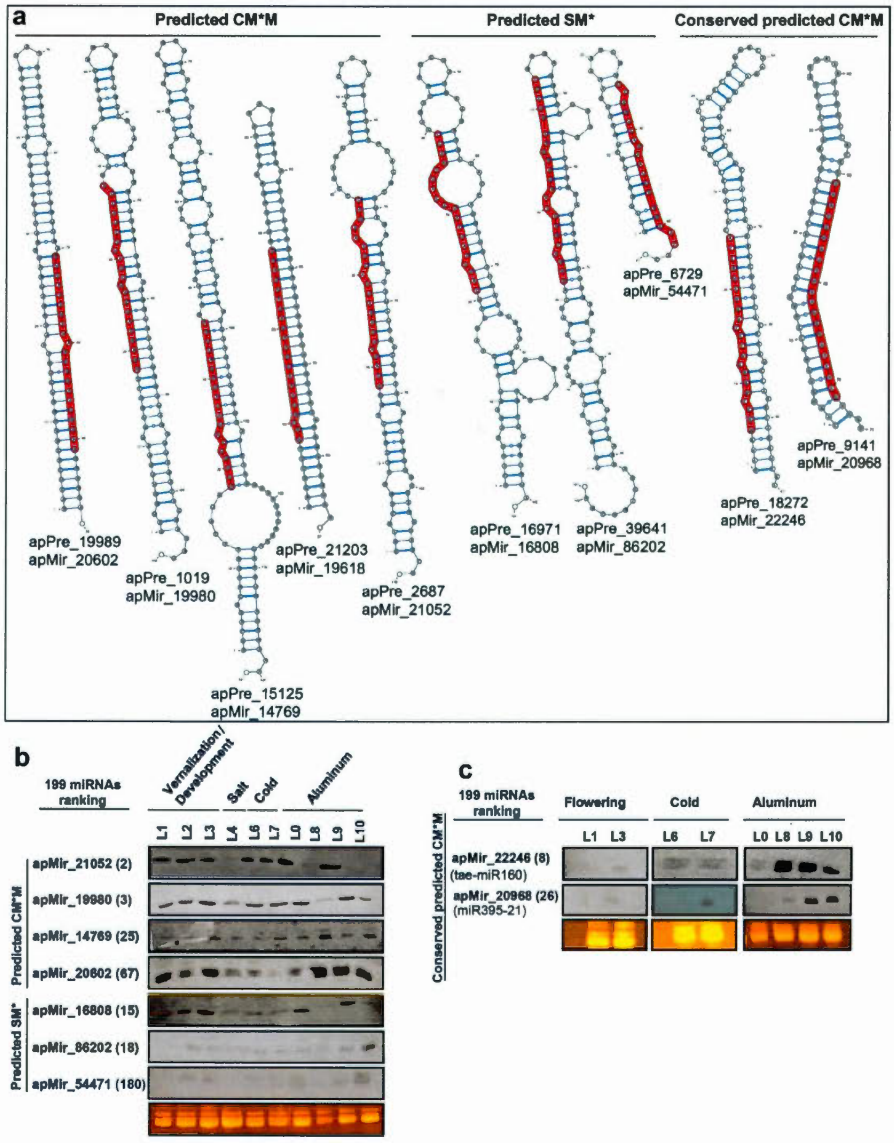


Figure 3.3: Experimental validation of predicted and conserved wheat miRNAs.

a) Pre-miRNA secondary structure of miRNA candidates experimentally validated by northern in the investigated libraries; **b)** northern blot of predicted miRNAs in common between MiRdup* and MIRcheck (CM*M) as well as specifically predicted with MiRdup* tool (SM*); **c)** northern blot of miRNA candidates identified by both sequence homology against miRBase (conserved miRNAs) and predicted in common between MIRcheck and MiRdup*. Ethidium bromide staining of the rRNAs is shown as gel loading control. L0 represents the control library for Al treatment (L8) in spring wheat Bounty which was not sequenced. The numbers between the parentheses correspond to the expression rank among the 199 predicted miRNAs. The lower value corresponds to the higher read abundance. For more information about the libraries and conditions see Method B.1.1 and Table B.3.

Table 3.2: Characteristics of selected miRNAs using MiRdup* and MIRcheck validated by northern blot

ID	MiRNAs	A	B	C	D	Associated conditions	Relevant targets (with uniref ID)	Training dataset
Predicted miRNAs in common with miRdup* and MIRcheck								
apMir_20602	GUCAUCUAUAUUGGAACGGAG	1	1	1	5	Salt, floral transition and flowering	Glutathione peroxidase (Q9SME6), putative phosphate phospho1 (M8CZ66)	BM
apMir_19980	AUAGCAUCAUCCAUCCUACCA	3	1	3	6	Al, floral transition and flowering	putative membrane-associated protein (gi|22548307|gb|BU100508.1|BU100508)	BPM
apMir_14769	GUUGUCAUAUAUGUAUUGA	2	1	2	6	Cold, al and salt	Putative RSH disease resistance-related protein (Q8H5X7), T-complex protein 1 subunit alpha (l3RZC6)	M

ID	MiRNAs	A	B	C	D	Associated conditions	Relevant targets (with uniref ID)	Training dataset
apMir_21052	UGAGAUGAGAUUACCCAAUAC	3	2	4	7	Cold, floral transition and flowering	NA	P
Predicted miRNAs specefic to miRdup*								
apMir_16808	CAUCGAUCAUCCAUCACCC	2	5	6	7	Not differentially expressed	Dehydrin, (CD909074, TA50415_4565), Phosphorylase (Q84P16)	B
apMir_86202	AGGGUCGGCCAGCGGUGCGGCCCGU	4	2,4	8	6	Cold, floral transition and al	NA	BM
apMir_54471	UCAGUCAUAAUCCGGCAC	3	1	3	7	Salt, al and floral transition	NA	MP
Conserved miRNAs predicted in common by miRdup* and MIRcheck								

ID	MiRNAs	A	B	C	D	Associated conditions	Relevant targets (with uniref ID)	Training dataset
apMir_22246 (tae-miR160)	UGCCUGGCUCCCUGUAUGCCA	3	1	3	9	Cold, salt, al, floral transition and flowering	Auxin response factor (M8BC98), Auxin responsive protein (R7WEP7)	BPM
apMir_20968 (miR395-21)	UGAAGUGUUUGGGGGAACUCU	2	1	2	8	Cold, salt, al, lowering	Bifuntional 3'-phophoadenosine 5' phosphosulfate synthase (M7ZFX2), ATP sulfurylase (M9T1P9)	BMP

The selection was based on several characteristics of the miRNA secondary structure in the duplex including, the number of bulges in the duplex (A), number of the successive unpaired bases in each bulge in the duplex (B), total number of the unpaired bases within the duplex (C), nucleotide number in the loop (D). Training datasets (B : all miRBase species ; P : all plants ; M : monocot only). The reverse complement sequences of miRNAs used as probes for northern blot validation are presented in Table B.13.

B.11). We also find that regulated miRNAs may exhibit either common or specific expression patterns. Many of them show expression that is tissue, stress, genotype, or development stage-specific (Figure 3.4c and d). They may be specific to Al in roots, cold/vernalization and salt treatment in aerial parts, or common to two stresses or to all of the investigated abiotic stresses (Figure 3.4c). This indicates a crosstalk between the regulatory mechanisms of cold, Al and salt responses. This observation is confirmed by northern blot analysis showing a dynamic and complex expression pattern for several abiotic stress responsive miRNAs (Figure 3.3b and c). For instance, the candidates apMir_19980 and apMir_16808 are slightly up-regulated by cold, but also strongly down-regulated by salt and Al. The regulated miRNAs may be also specific to vegetative (L1- L2), reproductive phase (L3) ; or common to the two phases (Figure 3.3b and Figure 3.4d). Moreover, out of the 199 miRNA candidates, less than 10% are ubiquitously expressed under the investigated conditions.

3.3.5 Functional classification of abiotic stress and developmentally regulated miRNAs in wheat

The potential functions of regulated miRNAs (differentially expressed between two conditions) were classified into 24 miRNA groups for cold (Co1-Co8), Al (Al1-Al8) and development (Dev1-Dev8) according to their expression in two wheat genotypes that differ in their degree of tolerance as well as during different development stages (Table B.12). For each stress, we found that groups 5 and 6 (Co5-Co6 for cold/vernalization ; Al5-Al6 for Al) having similar expression profiles in tolerant and sensitive genotypes are associated with cold/Al responses while other groups (Co1-Co4, Co7-Co8 and Al1-Al4, Al7-Al8) showing different expression patterns between the two genotypes are associated with tolerance (Table B.12). For development miRNA groups, 6 groups are associated with floral transition (Dev3-Dev4) and flowering (Dev1-Dev2 and Dev7-Dev8). The 24 miRNA groups

were subjected to GO enrichment analysis based on the 3 categories : cell component, molecular function and biological process (Figures B.4b-B.6b). Several highly enriched GO Slim terms associated with the studied conditions (Figure 3.5 and Figures B.4b-B.6b).

The miRNA group associated with cold responses (Co5) is specifically enriched for *membrane* (triose phosphate translocator) in the category cell component (Figure B.4b) and *signal transduction* (*Auxin-responsive proteins*) in the category biological process (Figure 3.5 and Figure B.6b). Consistently, enrichment is also found for nucleus (Figure B.4b), *protein binding activity* (Figure B.5b), *nucleobase containing component metabolic process and response to endogenous stimulus* (Figure 3.5 and Figure B.6b) which are all overrepresented by auxin responsive proteins. These results indicate that cold regulated miRNAs may function in carbon partitioning during photosynthesis and in auxin-activated signaling pathways. For miRNA groups associated with Al responses (Al5-Al6), an enrichment is found for *hydrolase activity* (protein phosphatase 2C, lipase), *catalytic activity* (glutathione peroxidase, phenylalanine ammonia-lyase) (Figure B.5b), *response to endogenous stimulus* (Auxin Response Factors), *DNA metabolic process* (histone 4) (Figure 3.5 and Figure B.6b). These results indicate that Al-regulated miRNAs may function in regulation of gene expression and signaling as well as plant defense under oxidative stress. More interestingly, many targets found for miRNA groups associated with cold (Co1, Co2, Co4) and Al (Al1, Al2 and Al3) tolerance are known for their function in stress adaptation. For miRNA groups associated with cold tolerance, the groups Co1 and Co2, showed enrichment for the GO Slim term *response to stress* (phosphoglycerate mutase, Defensin-like protein 1, Universal stress protein A-like protein). In addition, Co1 is enriched for *response to abiotic stimulus* (thaumatin-like protein, glutathione S-transferase) (Figure 3.5 and Figure B.6b) and the group Co2 is enriched for *cell wall* (Defensin-like protein, phospho-3-

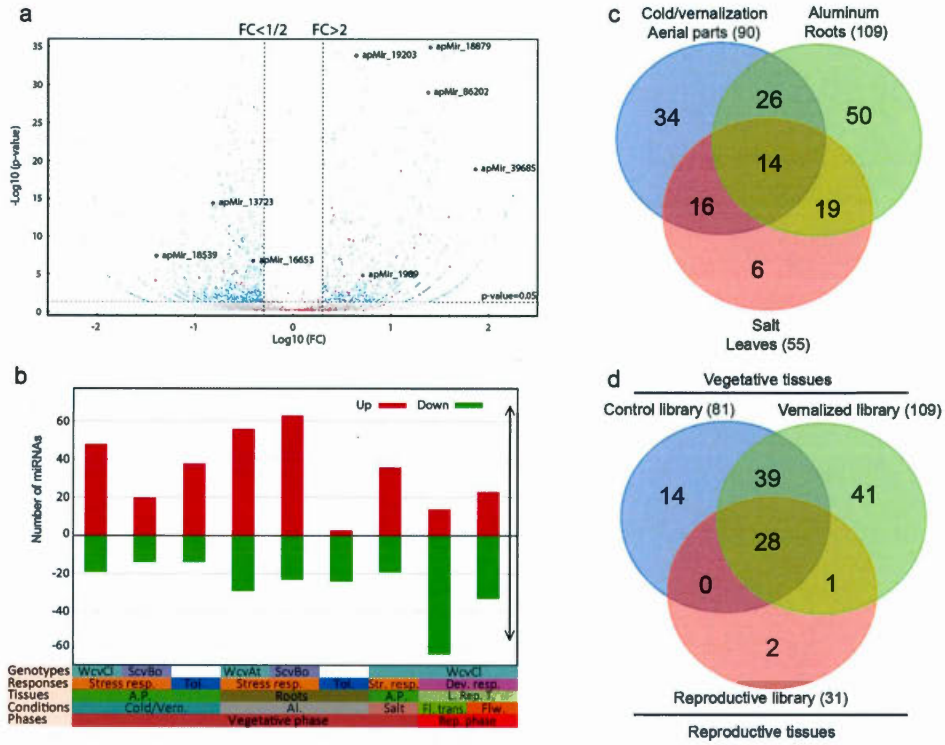


Figure 3.4: Differentially expressed miRNAs in response to cold, salt, aluminum and development. **a)** The differential expression of miRNAs in response to vernalization (presented on log10 adjusted p-value based on the FDR method of Benjamini and Hochberg (Hofacker, 2003), associated with the log10 of the fold change (FC)). The lines specify the thresholds used to identify the most relevant differentially expressed miRNAs. The blue and red dots correspond respectively to expressed small RNAs and predicted miRNAs; **b)** the frequencies of differentially expressed miRNAs in response to vernalization, cold, Al, salt and development stage (floral transition and flowering); and those differentially expressed between tolerant and sensitive genotypes; **c)** Venn diagram of miRNAs regulated under short/long exposure to cold (cold/vernalization, L2/L1 and L7/L6) in leaves, Al (L10/L9 and L8/L9) in roots and salt (L4/L1) in leaves; **d)** Venn diagram of miRNAs expressed in control plants during vegetative phase under normal conditions (control library L1), plants acclimated up to 56 days at 4 (vernalized library L2) during vegetative phase and, plants acclimated up to 56 days at 4 and then transferred to normal conditions under long day photoperiod to induce flowering during the reproductive phase (reproductive library L3). Up, up-regulated miRNAs; Dw, down-regulated miRNAs; Cold/vrn, cold and vernalization responsive miRNAs in spring (L7/L6) and winter wheat (L7/L6), respectively; salt responsive miRNAs in winter wheat (L4/L1); Al responsive miRNAs in spring (L8/L9) and winter (L10/L9) wheat; For tolerance, only differentially expressed miRNAs between cold (L2/L7) and Al (L10/L8) treated libraries are presented. All other abbreviations' are described in the legend of Figure 3.2. See Method B.1.1 and Table B.3 about libraries and conditions and Table B.11 for more information about regulated miRNAs.

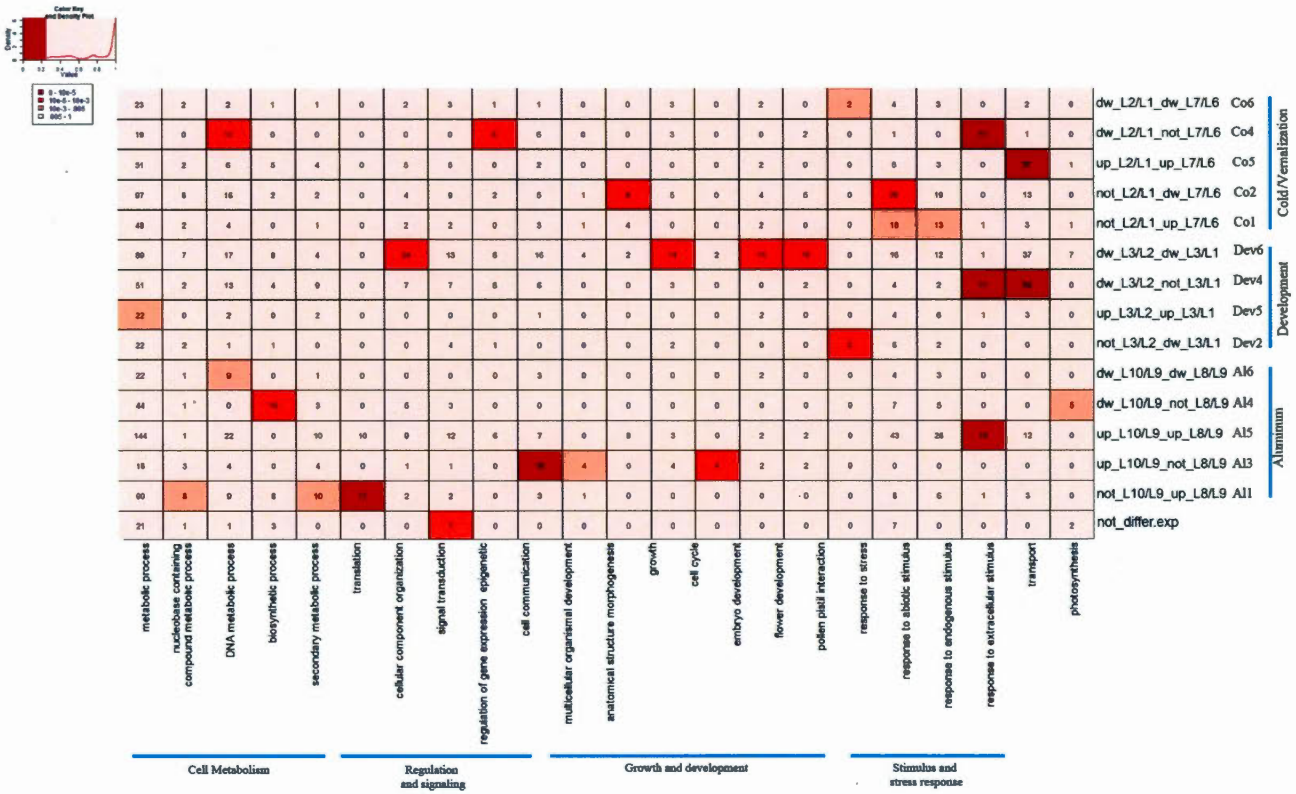


Figure 3.5: GO Slim enrichment for differentially expressed miRNAs in response to abiotic stress and development.

Differentially expressed miRNAs with the same or different expression patterns between plants from tolerant and sensitive genotypes under normal and abiotic stress conditions; and between plants at vegetative and reproductive phases were classified into 24 miRNA groups. MiRNA targets are annotated to the best scoring GO Slim terms in Biological process category. The lines are grouped according to their association to cold and vernalization (L2/L1 and L7/L6), Aluminum (L10/L9 and L8/L9) and development (L3/L1 and L3/L2). See Table B.12 for more information about miRNA groups. The value in each case indicates the number of associations miRNA-target- GO for the corresponding GO Slim. The enrichment is presented in four different colors ("brown square symbol" high enrichment (P-value $< 10^{-5}$), "orange square symbol" medium enrichment ($P - value < 10^{-3}$), "light orange square symbol" low enrichment ($P - value < 0.05$) and "white square symbol" no enrichment ($P - value = 0.05$)).

sulfolactate synthase-like), *nucleus* (CBFIVb-B20), *hydrolase activity* (Ubiquitin-like-specific protease, Serine carboxypeptidase) *and catalytic activity* (Gibberellin 20 oxidase, Glyceraldehyde-3-phosphate dehydrogenase) (Figure B.5b). This indicates that the identified cold regulated miRNAs may function in proteolysis, gibberellins biosynthesis and glucose metabolism. The group Co3 is enriched for *transporter activity* (Sodium/hydrogen exchanger) (Figure B.5b) while the group Co4 is enriched for *pollen-pistil interaction* m(Serine/threonine-protein kinase) (Figure 3.5 and Figure B.6b) indicating that they may also have a function in ion transport and signaling.

Moreover, for groups associated with Al tolerance, the most enriched GO Slim terms are *mitochondrion, transporter activity* (Cytochrome b-c1 complex) and *carbohydrates metabolic process* (glyceraldehyde-3-phosphate dehydrogenase) indicating that miRNAs from these groups may mediate glycolysis and respiration process under Al stress conditions. Enrichment is also found for the terms *catalytic activity, hydrolase activity, sequence specific transcription factor activity* (Figure B.5b), *secondary metabolic process, embryo development, anatomical structural morphogenesis* and *multi-organismal development* (Figure 3.5 and Figure B.6b). They are overrepresented by many important metabolism enzymes involved in phenylpropanoid metabolism and ubiquitination process (phenylalanine ammonia-lyase, DEAD-box ATP-dependent RNA helicase, Ubiquitin carboxyl-terminal hydrolase*)* as well as transcription factors (Homeobox-leucine zipper proteins).

Interestingly, the targets in miRNA groups associated with development are involved in cell growth and flowering. The miRNA groups associated with flowering (Dev1, Dev2) show enrichment for the GO Slim terms *protein binding activity, sequence specific DNA transcription binding activity* (MIKC-type MADS-box transcription factors), and/or *kinase activity* (Serine/threonine-protein kinase) (Figure B.5b). They are also enriched for *nucleobase containing compound metabolism*

and *pollen pistil interaction* that are specifically represented by two well characterized regulators genes (MIKC-type MADS-box and Serine/threonine-protein kinase) (Figure 3.5). One miRNA group associated with floral transition (Dev 4) is enriched for the GO Slim terms *transport* and *response to endogenous stimulus,* specifically represented by Auxin-responsive proteins. Furthermore, the potential function of ubiquitously expressed miRNA libraries was also investigated. They show significant enrichment for the GO Slim terms *thylakoid* (Figure B.4b), *kinase activity*, *nucleotide binding activity* (Figure B.5b), and *cell protein modifications* (Figure 3.5 and Figure B.6b). This result indicates that constitutively expressed miRNAs may modulate the basic cellular functions reflecting their vital regulatory role in other growth conditions yet to be identified in wheat.

3.4 Discussion and conclusions

3.4.1 The wheat miRNA pipeline

In this study, we developed a pipeline that identifies conserved as well as clade and species-specific or young miRNAs. This pipeline can be easily adapted for other plant species. To predict miRNAs from NGS and analyze their function, the steps described in Figure 3.1 are required. While several steps are standard in NGS analyses (Hsieh *et al.*, 2009; Jeong *et al.*, 2011; Ling *et al.*, 2011), we improved the miRNA prediction steps by integrating folded pre-miRNA candidates, expression profiling and functional analyses of differentially expressed candidates. To address the step of miRNA prediction, we decided to exploit two methods with different algorithmic schemes MiPred (Jiang *et al.*, 2007) and HHMMiR (Kadri *et al.*, 2009) to have a broad range of hairpin candidates. These methods were trained on pre-miRNAs from plants and wheat sequences available in miRBase and resulted in the identification of a large number of pre-miRNA candidates using the predictors MiPred (Jiang *et al.*, 2007) and HHMMiR (Kadri *et al.*,

2009). To address issues of latter methods for the lack of consideration of mature miRNA and their surrounding biological features, we developed a classifier that ranked the best 35 biological features of plant miRNAs that was integrated into MiRdup* (Table B.5). For robustness, the classifier's models were trained separately on three datasets (all miRBase species, all plants and only monocots). This increases species-specificity and allows the discovery of features that distinguish wheat miRNAs from those of other species. The developed classifier (MiRdup*) was able to reduce the level of false prediction obtained by MiPred (Jiang *et al.*, 2007) and HHMMiR (Kadri *et al.*, 2009) by more than 81% (Figure 3.1, Figures B.1 and B.2) and allowed the assessment of the position of a miRNA in a given pre-miRNA sequence. In addition, the combinatorial analysis between MiRdup* and MIRcheck (Jones-Rhoades et Bartel, 2004) which identifies 20-nt regions of a given plant pre-miRNA using a predetermined set of rules and constraints, show that MIRcheck is too stringent and easily removed experimentally validated miRNAs (Figure 3.3b and Figures B.2).

The availability of wheat EST databases and our approach enabled us to identify with confidence 199 miRNA candidates. These candidates may include miRNA gene homeologs from the three genomes of hexaploid wheat, or ESTs with SNP differences in different wheat varieties. It is also reasonable to assume that these families represent only a fraction of the total miRNAs that may exists in hexaploid wheat since many small RNAs still remain unmapped to wheat sequences or conserved miRNAs from miRBase. The availability of the complete assembled and well-annotated hexaploid wheat genome will help to complete the discovery of the remaining miRNAs.

It is important to emphasize that among the predicted miRNAs, in spite of being derived from ESTs, less than 5% of the mature miRNAs are associated with known protein coding regions and less than 7% are related to transposable elements

(Table B.7a and B.7b). According to Dinger *et al.* (2008), many transcripts are categorized as bifunctional RNAs. They can be translated into protein but also function independently as RNA. The presence of such bifunctional RNAs challenges the assumption that the RNA world can be neatly parsed between mutually exclusive protein-coding and non-coding categories.

3.4.2 MiRNA candidates associated with abiotic stress responses

This study represents one of the largest *de novo* miRNAome analyses in response to different abiotic stresses and development in hexaploid wheat. Although many cold responsive miRNAs have been identified in spring wheat using NGS (Tang *et al.*, 2012), our study identified a large number of novel candidates regulated by cold, vernalization, Al and salt with dynamic and complex expression patterns (Figures 3.4 b, 3.4 mb and Table B.11). Several identified miRNAs are either associated with a specific stress or common to at least two stresses (Figures 3.3 mb and 3.4 c). Many of their targets are known to be stress-related genes (Figure 3.5 and Figures B.4b-B.6b) commonly regulated under abiotic stresses.

Our results show that miRNAs may mediate plant responses to Al treatment by regulating expression of stress related genes particularly those involved in auxin signaling and fatty acid metabolism. This is consistent with the fact that Al affects the relative abundance of membrane lipids and the degree of fatty acid unsaturation (Wagatsuma *et al.*, 2005; Hossain Khan *et al.*, 2009) and Auxin Response Factors (*ARFs*) that are known to inhibit root development in response to Al toxicity (Wang *et al.*, 2005). In addition, the experimentally validated apMir_22246 (which corresponds to *tae-miR160*) is regulated by Al exposure (Figure 3.3 c) and targets specifically ARFs. Many ARF members are known to be regulated by *miR167 and miR160* and to play regulatory roles in adventitious rooting (Gutierrez *et al.*, 2012), supporting the possible role of apMir_22246 in root development

under Al treatment.

3.4.3 MiRNA candidates associated with cold responses and freezing tolerance

Our data indicate that cold regulates the expression of several miRNAs in spring as well as in winter wheat (Figures 3.3 b, 3.4 b and c). Four miRNA groups associated with cold tolerance (Table B.12) target a set of cold regulated genes known to be involved in freezing tolerance including the transcription factors *CBFs*, dehydrins, DEAD-box RNA helicases, thaumatin-like protein (Badawi *et al.*, 2008; Kim *et al.*, 2008; Kurepin *et al.*, 2013; Janska *et al.*, 2010). Interestingly, many candidate miRNA target genes related to the ICE1–CBF major pathway that regulates freezing tolerance in cold hardy plants. This includes the targets DEAD-box ATP-dependent RNA helicase 12, CBF and dehydrin (Table B.9). Results from our previous studies demonstrated that genes related to the ICE1–CBF pathway play a critical role in freezing tolerance in hexaploid wheat (Badawi *et al.*, 2007). Here we show that the miRNA candidate apMir_16808 is regulated in response to cold (Figure 3.3 b), and target the cold responsive genes dehydrins (Table Tab2) (Badawi *et al.*, 2008; Janska *et al.*, 2010). The candidate apMir_19532 from miRNA group associated with cold tolerance target CBFIVb-B20 gene (Table B.9). These results suggest that these miRNAs may contribute to freezing tolerance by regulating cold-regulated genes belonging to the CBF regulon in winter wheat.

3.4.4 Predicted miRNA target genes common in regulating several stresses

Plants evolved common regulatory mechanisms to adapt to environmental stresses such as oxidative stress commonly induced by both cold and Al. Our results show that many of the identified abiotic stress responsive miRNAs exhibited a common stress expression pattern (Figures 3.3 b and c and 3.4 c). For instance, the expression of the new member of miR395 family, miR395-21 corresponding to

apMir_20968, is commonly regulated in response to cold and Al stress (Figure 3.3 c) indicating that miR395 is not specific to sulfate starvation as previously reported in *Arabidopsis* and rice (Jeong *et al.*, 2011; Liang et Yu, 2010). Zhao *et al.* (2013) also reported that miR395 is involved in phosphate homeostasis in wheat. This indicates that miR395 mediates not only plant response to sulfate deficiency but also may mediate responses to other nutrients that are imbalanced under abiotic stress conditions. Taken together, our results indicate that miR395 would play a common role in plant nutrient homeostasis under abiotic stress conditions. In agreement with previous suggestions, our results indicate that miRNAs coordinate crosstalk among different nutrient deficiencies. This is the first indication that crosstalk between cold, Al stress and plant nutrients could be regulated by miRNAs. Moreover, we show that the miRNA candidate apMir_20602 is also commonly regulated under cold, salt and Al (Figure 3.4 b) and targets glutathione peroxidase (Table Tab2). Recent findings showed that human miRNAs regulate glutathione peroxidase expression to maintain redox homeostasis (Wang *et al.*, 2014). This supports the possible role of apMir_20602 in mediating crosstalk between abiotic stress responses by regulating glutathione metabolism.

3.4.5 Wheat vernalization responsive miRNAs associated with floral transition and flowering

In this study, we investigate the role of miRNAs during the transition from the vegetative to the reproductive phase, and during flowering in winter wheat that requires vernalization to flower. We found that among developmentally responsive miRNAs, many candidates target cold responsive genes known for their function in flowering transition and flower development (Table B.9). For instance, the candidate apMir_19892 corresponding to hvu-miR444b (Data B.2.1) could target many MIKC-type MADS-box transcription factors, the homologs of *TaAGL17* and *OsMADS57*. In wheat, MIKC-type MADS-box transcription factors control

flower development and morphogenesis (Paolacci *et al.*, 2007). In barley, this target contains both the target site for miR444b and the precursor sequence for miR444a (Colaiacovo *et al.*, 2012)). In rice, *OsMADS57* is involved in axillary bud development and regulation of tillering through down-regulation of miR444a (Guo *et al.*, 2013). Since the miRNA variants from miR444 family are functional, and *MADS*-box genes are collectively regulated by the miR444 family (Sunkar *et al.*, 2012), we suggest that apMir_19892 may mediate flowering through the regulation of MIKC-type MADS-box transcription factor gene expression. ApMir_19532 target genes encoding Ubiquitin-like-specific protease *ESD4* known to regulate plant responses to cold and the time of flower initiation (Reeves *et al.*, 2002; Murtas *et al.*, 2003). In addition, apMir_20860 corresponding to miR159 (Data B.2.1) are involved in promotion of floral transition in many species. In ornamental plants, miR159-regulated *GAMYB* expression is an effective pathway of flowering time control Li *et al.* (2013a). This suggests that apMir_19532 and apMir_20860 may mediate flowering time in wheat through the regulation of Ubiquitin-like-specific protease ESD4 and *GAMYB* gene expression.

3.5 Methods

3.5.1 Plant material and small RNAs isolation

In this study, three genotypes of hexaploid wheat (*Triticum aestivum L.* 2n=6x=42, AABBDD), one spring genotype (cv Bounty, cold and Al sensitive) and two winter genotypes (cv Clair, cold tolerant and Atlas66, Al tolerant genotype) were used to construct ten different small RNA libraries from plants in vegetative and/or reproductive stages and/or exposed to different stress treatments or under normal conditions (Method B.1.1 and Table B.3). To identify miRNAs that are associated with different development stages, tissues from both vegetative and reproductive phases were used. Vegetative phase samples include leaves and crown from the

aerial part of plants. Reproductive phase samples include leaves at flag leaf stage, developing spikes with sizes ranging from 2 to 110 mm, and spikes partially and completely-opened with and without pollen. We used also root tips to identify miRNAs associated with Al and salt stress, and aerial parts including leaves and crown to identify miRNAs associated with cold and salt. In addition, we used different genotypes of winter (tolerant) and spring (sensitive) wheat to identify miRNAs associated with Al and freezing tolerance. Small RNA extraction was initiated from 200 mg of a mixture of leaves, stem or root tip tissues from 10 to 100 seedlings for each time point. Control and treated plants were sampled at the same time of the day for each time point (except for the first day where a few samples were taken at short time points) as described in (Method B.1.1 and Table B.3). Small RNAs (below 200 nt) were isolated from each sample using the mirVana miRNA Isolation Kit (Ambion Inc. US). MiRNAs (small RNAs below 40 nt) from each time point were isolated using the flashPAGE fractionation kit (Ambion) (Method B.1.1 and Table B.3), and then purified using the flashPAGE Reaction Clean-up kit (Ambion) according to the manufacturer's protocols. Their integrity was assessed using a DNA 1000 LabChip on an Agilent 2100 Bioanalyzer (Santa Clara, CA, USA).

3.5.2 MiRNA libraries construction and sequencing

Twenty five nanograms of purified miRNAs from each time point of a given condition (Method B.1.1 and Table B.3) were pooled and used as a template to produce the corresponding miRNA library. MiRNAs were tagged with a barcode system containing ten unique and specific amplification primers (1 barcode/library) and ten cDNA libraries were produced using the SREK kit (small RNA expression Kit, Ambion) according to the manufacturer's protocol. The libraries were sequenced on the SOLiD Analyzer according to the standard protocol (V2.1 Applied Biosystems).

3.5.3 Experimental validation of predicted miRNAs

For each library, identical amounts of plant tissues from each time point were ground and mixed with TRIzol Reagent (Life Technologies). The same extract volume from each time point of each library was pooled to isolate small RNAs using the mirVana miRNA Isolation Kit. Five micrograms of small RNAs from each library were analyzed by northern blot (Várallyay *et al.*, 2008). The experiment was repeated at least twice for each selected probe. The oligonucleotide probes are presented in Table B.13.

3.5.4 Identification and extraction of potential pre-miRNA candidates from sequenced small RNAs

From the 89 million reads obtained from the ten libraries, we first removed reads that have low quality scores as recommended for SOLID sequencing (Hackenberg *et al.*, 2009; Ribeiro-dos Santos *et al.*, 2010) (Table B.4). After adapter removal using the program cutadapt v0.9 (Schulte *et al.*, 2010), small RNAs between 18 and 30 nt were mapped to ESTs with the MAQ v07.1 program (Ondov *et al.*, 2008) allowing a maximum of two mismatches (Method B.1.3). Then, sequences with low complexity or containing repeats were filtered out using RepeatMasker v3.2.9 (Smit *et al.*, 2010) with RepBase15.09 and Repeatmasker-Libraries-20130422 (Jurka *et al.*, 2005), and a slow search method against *Triticum aestivum* and *Oryza sativa*. For each mapped EST, we considered two sequences that could include a potential pre-miRNA candidate as follows : 20 nt before the start and 160 nt after the end of the mapping were extracted from both EST strands.

The secondary structures of the extracted sequences (pri-miRNA) were folded with RNAfold from ViennaRNA v1.8.4 package (Hofacker, 2003) to identify those having a hairpin-like shape, one of the fundamental characteristics of pre-miRNAs. Then, these sequences were submitted to two pre-miRNA predictors using different

algorithmic schemes, HHMMiR (Kadri *et al.*, 2009) and MiPred (Jiang *et al.*, 2007), which were trained with pre-miRNAs of all cloned or sequenced miRNAs from miRBase (Kozomara et Griffiths-Jones, 2011). Finally, to identify conserved miRNAs, we performed a *blastn* of small RNAs with more than 100 reads in at least one library against miRBase (V21) (Kozomara et Griffiths-Jones, 2014) with a *word_size* 7, *maximum e-value* 0.1, *percentage identity* 80, *gap open penalty* 5, *gap extension penalty* 2, *match score* 1, *mismatch score* -2, *filter low complexity.* We restricted the blast results as follows : query coverage of 90%, subject coverage of 90%, with no gap allowed.

3.5.5 Filtering false positive pre-miRNAs

We adapted our previously published machine learning classifier (Leclercq *et al.*, 2013) to the best plant features associated with the position of miRNA-miRNA* duplex in the pre-miRNA (Method B.1.4), and we named this version MiRdup*. The latter differs from the original one on the 35 retained features (Table B.5) that are relevant for plants and has been trained on all experimental (cloned or sequenced) miRNAs from miRBase subdivided in three datasets (all species, all plants and only monocots). This classifier computes a score of prediction scaled between 0 and 100 (more evidence). The potential pre-miRNAs from MiPred or HHMMiR that obtained a MiRdup* classification score higher than 90 and with miRNA read abundance above 100 in at least one library were selected as candidates. In addition, to help identifying the potential functional miRNAs among several candidates, we applied the relaxed expression rules derived from the update of the specific criteria for plant miRNA annotation reported by Meyers *et al.* (2008). In parallel, we applied MIRcheck (Jones-Rhoades et Bartel, 2004), a well-known tool for plant miRNA identification, on the overall predicted miRNAs to compare the differences with MiRdup*. A combinatorial analysis between the two tools is provided (Figures B.1 and B.2). Then, the pre-miRNAs were blasted

against TREP database to identify miRNAs that originate from transposable elements (Table B.7a). The e-value threshold used is 5.0E-05 and Hit Coverage $(HC) = 85$ and percentage of identity $= 80$. We performed also a blastx of ESTs producing the identified pre-miRNAs against proteins coding sequences from protein plant database with default parameters (Table B.7b). The threshold to include a blast hit of an EST into a given protein core is an e-value lower than $1.00E-20$ and query coverage or hit coverage higher than 85% and percentage of identity higher than 75%.

3.5.6 Statistical analyses of the abundance of potential miRNAs

To quantify and compare sequence abundance across different libraries, raw read counts were normalized using rpm (reads per million). Sequences with read counts lower than 100 in all libraries are removed. Significance level of the difference of small RNA between two libraries was analyzed using a corrected Z–Score method as described in Kal *et al.* (1999). An adjustment for multiple comparisons based on the false discovery rate (FDR) (Benjamini et Hochberg, 1995) was performed $(FDR < 5\%)$. The small RNAs with fold change lower than 0.5 or higher than 2.0 were retained.

3.5.7 MiRNA target analyses and GO enrichments

MiRNA target genes were identified using the FASTA engine of Tapir v1.0 program, with a stringent maximum score of three and minimum free energy = 0.7 (Bonnet *et al.*, 2010) excluding ESTs annotated as unknown proteins. For the obtained and annotated target genes, we retrieved their classification in Gene Ontology (GO) through the GO Slim viewer on AgBase webserver (Binns *et al.*, 2009). The GO Slim enrichments were performed using the standard hypergeometric test. The wheat genome GO Slim background was constructed taking into

account the overall GO Slims covered by the 127,039 UniRefs id retrieved from all the collected ESTs Database. The GO Slim terms with $P - value < 0.05$ were considered as enriched. The same procedure was applied for the targets of the differentially expressed miRNAs. Unlike the overall analysis, the GO Slim background considered in each condition was computed from only the GO Slim of the identified target genes present in the two compared libraries from sensitive and tolerant genotypes. For functional analysis, we investigate the potential function of all identified miRNAs in the 10 investigated conditions (10 libraries) for miRNAs having at least 100 reads in at least one sequenced library. (data B.2.3).

3.5.8 Availability of supporting data

All the predicted and conserved miRNAs from this study, the published miRNAs from the literature, the small RNA expression profiles are provided at the following database `http://wheat.bioinfo.uqam.ca`.

3.6 Declarations

Acknowledgement

We would like to thank Mathieu Blanchette for his useful comments and suggestions. Thanks also to all people and institutions who gave us access to their super computers, Alix Boc for the *Trex cluster*, Daniel Lemire in the Licef research center for the cluster Erasme at the TeluQ, the Clumeq (Supercomputer Consortium Laval UQAM McGill and Eastern Quebec) for the clusters Colosse and Guillimin, as well as HHMMiR and Tapir authors for providing source code for this project.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants to FS, MH and ABD ; NSERC fellowship for EL and MAR, as well as the Fonds de recherche du Québec-Nature et technologie (FRQNT) to ML and MAR.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZA, MH, ABD and FS designed the overall study. ZA performed the biological experiments and the construction of the libraries. MB and ZA performed the experimental validations. ML, MAR, EL, ABD designed the bioinformatics pipelines, data integrations and result aggregations. ZA, ML, MAR, ABD, FS analyzed and interpreted the results. ZA, ML, MAR, MB, JD, MH, ABD and FS contributed to the manuscript preparation. All authors read and approved the final manuscript.

Copyright

CHAPITRE IV

A NOVEL COMPREHENSIVE WHEAT MIRNA DATABASE, INCLUDING RELATED BIOINFORMATICS SOFTWARE

Auteurs	Mohamed Amine Remita, Etienne Lord, Zahra Agharbaoui, Mickael Leclercq, Mohamed A Badawi, Fathey Sarhan and Abdoulaye Baniré Diallo
Journal	Current Plant Biology
Édition spéciale	Genomic resources and databases
État de l'article	publié le 8 novembre 2016

4.1 Abstract

MicroRNAs (miRNAs) are emerging as important post-transcriptional regulators that may regulate key genes responsible for agronomic traits such as grain yield and stress tolerance. Several studies identified species and clades specific miRNA families associated with plant stress regulated genes. Here, we propose a novel resource that provides data related to the expression of abiotic stress responsive miRNAs in wheat, one of the most important staple food crops. This database allows the query of small RNA libraries, including *in silico* predicted wheat miRNA sequences and the expression profiles of small RNAs identified from those libraries. Our database also provides a direct access to online miRNA prediction software

tuned to *de novo* miRNA detection in wheat, in monocotyledon clades, as well as in other plant species. These data and software will facilitate multiple comparative analyses and reproducible studies on small RNAs and miRNA families in plants. Our web portal is available at : http ://wheat.bioinfo.uqam.ca.

Keywords : Abiotic stresses and development Expressed Sequenced Tags miRNA database Next-Generation Sequencing Wheat

4.2 Introduction

Next-generation sequencing offers interesting tools for generating and searching for microRNAs (miRNAs) that are expressed during growth and development. This approach provides several small RNA libraries and miRNA candidates from hexaploid wheat (*Triticum aestivum L.*) grown under different stress conditions such as heat, cold, and powdery mildew and *fusarium* infection. MiRBase, the current general database of miRNA sequences, is the main resource to access miRNAs data. Unfortunately, only 119 experimentally validated wheat miRNAs are available in the latest release of miRBase (Kozomara et Griffiths-Jones, 2011). Furthermore, the miRBase interface is not suitable for the analysis and assessment of miRNA candidates according to their original biological libraries and experimental conditions. To circumvent this limitation, we developed a novel comprehensive web-based server, WMP (Wheat MiRNA web-Portal). It is dedicated to wheat miRNAs, with emphasis on stress responsive miRNAs from different wheat genotypes and tissues. This database stores and displays differential gene expression data from several small RNA libraries. It also contains the data from 20 different miRNA studies (references are listed in the database interface). Moreover, the web portal yields and integrates views of the pre-miRNA hairpin structures and the related predicted targets. It also includes two miRNA predictors : *miRdup* (Leclercq *et al.*, 2013) and *MIRcheck* (Jones-Rhoades *et al.*, 2006) tuned for different

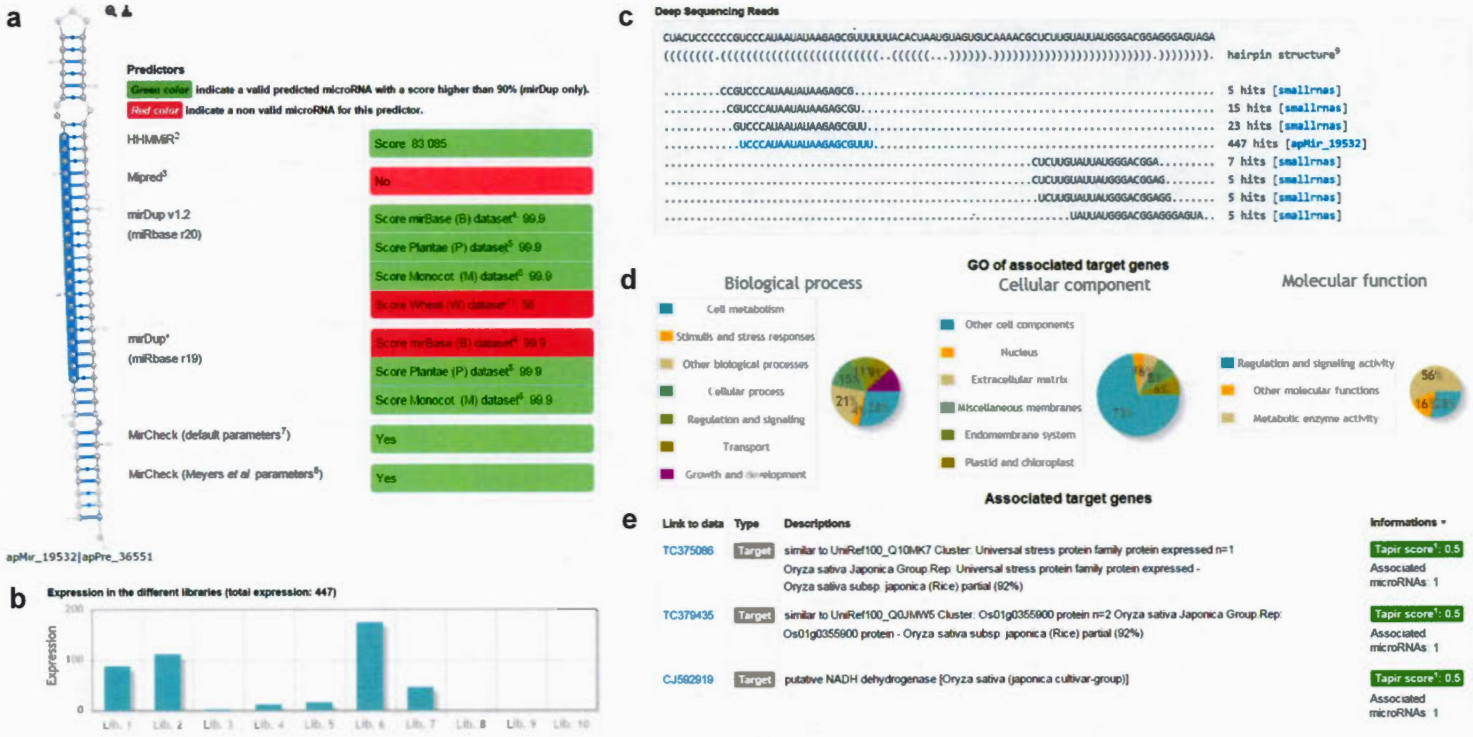


Figure 4.1: Overview of the main features of our miRNA database including (**a**) two-dimensional hairpin structures and the associated prediction scores for the miRNA apMir_19532, (**b**) expression view of each library, (**c**) view of small RNA expression and miRNA position against the computed pre-miRNA, (**d**) Gene Ontology annotations associated with a set of target genes and (**e**) description of the obtained target genes.

evolutionary clades (wheat, cereals or plants).

4.3 Database content and statistics

The current database includes data of ten small RNA libraries produced from plants grown under different abiotic stress conditions and development stages. From these libraries, a set of 168,834 unique expressed small RNAs are illustrated as well as 267 evolutionary conserved miRNAs according to miRBase (Kozomara et Griffiths-Jones, 2011). The database also contains a broad description of 5,036 published wheat miRNAs. It provides detailed presentation of 199 newly identified wheat miRNAs, 1,390 associated target genes, and 1.4 millions ESTs with 127,039 Uniref clusters, collected from seven wheat databases issued from Agharbaoui *et al.* (2015). Putative target genes were identified for each miRNA using the Tapir program (Bonnet *et al.*, 2010). Gene ontology (GO) associations and enrichments were also identified for associated targets resulting in 2,561 biological process, 1,616 cellular component, and 1,386 molecular function associations. To enhance the visualization of miRNA structures, a high quality picture of each miRNA and each pre-miRNA hairpin were generated using the Varna package (Darty *et al.*, 2009). Figure 4.1 highlights an example of graphical and statistical features of miRNA candidates in the database.

4.4 User interface

The database provides the following four main options : A) *Basic search* : allows keywords corresponding to one or multiple entries from the following categories : validated miRNAs, pre-miRNAs or the hairpin-folded structure motifs (dotbracket notation), associated target gene accession identifier, Uniref identifier or name, EST sequences or identifications, and GO descriptions or identifiers (e.g. GO :0008152). B) *Advanced search* : the user can search for differentially expressed

miRNAs by choosing the growth conditions and selecting a list of miRNA patterns (upregulated, downregulated, not differentially expressed, or not related to any of the conditions). The metrics for statistical significance (*p-values*) and the miRNA expression fold-change between libraries could also be set to filter candidates. C) *Data* menu : provides a direct access to predicted miRNAs, conserved miRNAs, the associated target genes and ESTs. This menu also gives access to *Libraries and conditions* option. This option allows profiling of all miRNAs expressed in any given libraries or under any given stress conditions. D) *Tools* menu : provides five important applications for studying small RNAs libraries and miRNAs. It includes *MIRcheck v1.0* (Jones-Rhoades et Bartel, 2004), *miRdup v1.2* (Leclercq *et al.*, 2013), *Small RNA finder*, *Library comparison* as well as a *Blast* interface (Altschul *et al.*, 1990; Camacho *et al.*, 2009). *MIRcheck* and *miRdup* are miRNA prediction tools. For both predictors, users can submit candidate miRNAs, miRNA precursors, or hairpin secondary structures for analysis. The *MIRcheck* tool allows two modes of computation (default parameters used by Jones-Rhoades et Bartel (2004) and the universal plant miRNA criteria given by Meyers *et al.* (2008)). The *miRdup* predictor (based on a machine learning approach) allows a selection among four evolutionary clade-training sets. These sets are extracted from experimentally validated miRNA sequences found in miRBase (all miRNAs of miRBase, viridiplantae (plant), monocotyledons and wheat (*Triticum aestivum L.*)). *Small RNA finder* option permits users to search for multiple sequences in a single query against putative miRNA and small RNA database. The *Library comparison* option allows extracting differential expressed small RNAs between two libraries. In the *Blast* option, users can identify the location of query sequences (nucleotide and amino acid sequences) against the wheat whole sequenced genome using the *Blast* tool. Further details on our database and a demonstration of some of its features are provided in the following section.

4.5 Study case : searching for miRNAs regulating glutathione S-transferases

The usefulness of our miRNA database is shown with an example of the search of miRNAs that regulate glutathione S-transferase (GST) enzyme family in wheat. GSTs are multifunctional proteins known for their important roles in both normal cellular metabolisms and the detoxification of a wide variety of products under stress conditions (Marrs, 1996; Jain *et al.*, 2010). GSTs have been shown to be miRNA targets in various other plant species exposed to stresses, such as osa-miR1848 in rice (Li *et al.*, 2010; Shaik et Ramakrishna, 2012), rsa-miR156 in radish (Xu *et al.*, 2013), mir168 homolog in sweet potato (Dehury *et al.*, 2013) and ssp-mir169 in sugarcane (Gentile *et al.*, 2013; Menossi *et al.*, 2015). Using the basic keyword search option, we first looked for the complete name of the GST enzyme : "glutathione S-transferase". We found 31 target gene sequences, one miRNA gene and 30 gene ontology annotations. To ensure that we didn't miss any other related data in the database we searched for alternative keywords related to the enzyme such as glutathione transferase and GST. Clicking on the target sequence identifier CJ893705 (with a good Tapir score of one), allows the user to access further related information such as the alignment with the miRNA apMir_11506, the Uniref clusters where the target belongs and the associated GO annotations. Information related to the found miRNA such as its sequence, length and expression details, can be easily obtained by clicking on its identifier link. In the *in silico* evidence section of the miRNA page, one could find that it is embedded in two precursors (apPre_12405 and apPre_8495) and it is predicted as valid miRNA with two different predictors (*miRdup* (Leclercq *et al.*, 2013) and *MIRcheck* (Jones-Rhoades et Bartel, 2004)). A deep sequencing reads view of precursors and their mapping small RNAs is appended to the evidence section. Interestingly, we noticed that apMir_11506 is regulated by cold and salinity. Furthermore, in miRBase section, we found that it has a 20-nucleotides

homolog miRNA of Arabidopsis thaliana (ath-miR8175). These results show that in total, 5 miRNAs are associated with GST : apMir_54471, apMir_15117, apMir_18589, apMir_11506 and apMir_19203. The first four miRNAs target the GST enzyme. apMir_18589 and apMir_19203 are generated by ESTs associated with GST Uniref. Analyzing the expression behavior in the different conditions reveals that miRNAs targeting the GST were regulated in aluminum, salt and cold stresses from sensitive and tolerant wheat genotypes.

4.6 Conclusion

The WMP database presents a novel resource for the analysis of miRNAs in cereals. This first version offers an access to miRNAs and small RNAs in the context of their association with different growth and development stages under stress conditions in wheat. It constitutes a unique entry for wheat expressed small RNAs and miRNAs for several studies. Furthermore, it provides a direct access to useful miRNA predictors for wheat species, as well as for other cereals and plants. In the future, we plan to update this database regularly by adding new libraries prepared by our team for wheat grown under different stress conditions and newly published small RNAs in wheat and other cereals. We will also includes an evolutionary toolkit to study miRNAs within all cereals. Further development will include option to integrate complex miRNA analysis pipelines using the Armadillo workflow platform (Lord *et al.*, 2012).

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) to FS and ABD and the Fonds de recherche du Québec - Nature et technologies (FRQNT) to ABD. MAR, EL and ML are FRQNT fellows. MAR is a NSERC fellow.

CONCLUSION ET PERSPECTIVES

Dans ce mémoire nous avons abordé l'utilisation des approches d'apprentissage automatique dans l'intégration des données biologiques, la construction des plateformes d'analyses et la classification de séquences biologiques. Ce travail de recherche a abouti à la rédaction de trois articles scientifiques.

Le premier article (chapitre 2) démontre l'efficacité de l'utilisation de l'apprentissage supervisé pour la classification des séquences virales en rangs taxonomiques. L'étude décrite dans cet article révèle multiples contributions : une nouvelle méthode de calcul d'attributs à partir des séquences, la généricité de l'approche de classification et une plateforme web intégrative et automatique pour la classification des séquences (CASTOR). Le calcul des attributs (*features*) est fait par un algorithme inspiré de la technique de biologie moléculaire (RFLP). Le RFLP *in vitro* est utilisé avec succès, pendant des décennies, dans l'identification des types viraux. L'algorithme RFLP *in silico* calcule des attributs décrivant la distribution des sites de restriction le long des séquences génomiques. Il a montré une certaine performance dans la discrimination de plusieurs types de virus (VPH, VHB et VIH). Cet algorithme est indépendant de la structure et de la fonction des séquences génomiques, ce qui confère une des qualités importantes de l'approche qui est la généricité. Ensuite, des modèles d'apprentissage sont entrainés sur les données virales étiquetées et utilisés pour la prédiction. La nouvelle approche peut être utilisée pour la classification d'un grand nombre de familles de virus dont leurs classifications dépendent sur les séquences de leurs gènes ou de leurs génomes et non sur d'autres caractéristiques tels que la morphologie ou la

virulence. Elle est générique et compétitive avec des outils spécifiques pour la classification des virus (COMET (Struck *et al.*, 2014) et REGA (de Oliveira *et al.*, 2005)). Finalement, l'approche a été implémentée dans une plateforme web publique, appelée CASTOR, dédiée à la classification des séquences nucléotidiques. CASTOR offre plusieurs applications et modes pour la construction des modèles d'apprentissage et la prédiction des classes. La plateforme permet le téléchargement des modèles et leurs partages dans une base de données ; ce qui permet aux pairs la réutilisation et la reproduction des expériences de classification dans un environnement unique et avec ou non les mêmes configurations.

Le deuxième article (chapitre 3) décrit une nouvelle méthode pour l'identification et la classification des miARNs à partir des données de séquençage des ARNs (RNA-seq). Contrairement à la méthode de classification décrite dans le chapitre 2, cette méthode combine plusieurs étapes de traitement de données. L'aspect intégratif de la méthode est dû à la complexité et l'immensité des données biologiques utilisées (des millions de petites séquences d'ARNs et de séquences ESTs, un génome hexaploïde partiellement séquencé, plusieurs conditions biologiques, etc.). Cette méthode a été développée à partir des données du blé mais elle peut être facilement adaptée à d'autres espèces (animales et végétales). Le pipeline conçu intègre des méthodes de classification basées sur la conservation (Blast (Altschul *et al.*, 1990)), des méthodes basées sur des filtres (MIRcheck (Jones-Rhoades et Bartel, 2004)) ainsi des méthodes basées sur l'apprentissage supervisé (MiPred (Jiang *et al.*, 2007), HHMMiR (Kadri *et al.*, 2009) et MiRdup* (Leclercq *et al.*, 2013; Agharbaoui *et al.*, 2015)) et l'apprentissage non supervisé (clustering). MiPred et HHMMiR ont été utilisés pour l'identification des précurseurs des miARNs candidats à partir des séquences qui se replient en épingle à cheveux (*hairpin*). Plusieurs régions génomiques se replient en *hairpins* et ressemblent à des vrais précurseurs de miARNs ce qui induit la génération de faux positifs par ces deux

prédicteurs. Afin de pallier ce défaut, un modèle de MiRdup (Leclercq *et al.*, 2013) est entrainé avec les 35 meilleurs caractéristiques biologiques des miARNs et avec trois lots de miARNs de différents niveaux taxonomiques (MiRdup*). Le pipeline a permis l'identification de nouveaux miARNs spécifiques à la plante et des miARNs conservés dans d'autres espèces, ainsi la classification fonctionnelle des gènes ciblés par ces miARNs.

Le dernier article (chapitre 4) présente le Wheat MiRNA web-Portal (WMP), une nouvelle plateforme web dédiée à l'analyse des miARNs du blé. La base de données de la plateforme dispose de plusieurs librairies de petits ARNs et de miARNs du blé impliqués dans le développement, la réponse et la tolérance de la plante aux conditions de stresses abiotiques (froid, salinité et aluminium). Une vue entière est consacrée pour chaque miARN détaillant ses caractéristiques de biogenèse, de conservation et fonctionnelles. La plateforme intègre des outils de classification utilisés et développés dans le chapitre 3.

Les approches de classification des séquences nucléotidiques basées sur l'apprentissage automatique ont montré des qualités supérieures (rapidité, frugalité et accessibilité) et des performances compétitives par rapport aux autres méthodes. Cependant, le développement des techniques liées à l'apprentissage automatique rencontre des difficultés et des limites. Ces limites peuvent être dues à la complexité des données traitées, la problématique de la classification des données biologiques et/ou aux approches d'apprentissage automatique utilisées.

La description des séquences par des attributs ou des caractéristiques est importante dans le processus de discrimination entre les classes. Une description des données par des types d'attributs adéquats et pertinents maximise une cohésion intra-classe et une séparabilité inter-classe. Les indices de silhouette (Rousseeuw, 1987) et de cohésion (Daigle *et al.*, 2015) calculés à partir des attributs RFLP

in silico sont plus faibles pour la classification du VIH-1 par rapport à celles du VPH et VHB. Les séquences des sous-types du VIH-1 sont plus divergentes, ce qui rend difficile aux attributs RFLP de les discriminer. Bien que CASTOR obtient des performances supérieures à 90% pour le VIH-1, le développement d'une nouvelle métrique RFLP *in silico* plus élaborée que *CUT* et *RMS* (tenir compte des positions des coupures par exemple) est nécessaire pour maximiser la cohésion interne et la séparabilité externe des classes. Ces deux dernières qualités aident à la construction d'un modèle d'apprentissage simple et à éviter des situations de surapprentissage (*overfitting*).

Comme discuté dans le chapitre 2, CASTOR a un taux de faux positifs (TFP) plus élevé que ceux de COMET (Struck *et al.*, 2014) et REGA (Alcantara *et al.*, 2009) dans la classification des séquences de VIH-1. Plusieurs approches (combinées ou non) peuvent être proposées pour résoudre ce problème. L'ajout d'une classe négative est une approche que nous sommes en train d'évaluer. Les pseudo-séquences virales sont générées à partir des vraies séquences virales avec des rétrécissements, des duplications et des réarrangements aléatoires. Une autre approche serait la redéfinition de la problématique de classification dans un espace de classes ouvert(*open-set classification*). Cette approche réduirait le TFP et accorderait aussi à CASTOR une robustesse dans la prédiction des séquences appartenant à des classes inconnues. Enfin, la méthode de classification peut calculer un score ou une probabilité d'inclusion ou d'appartenance d'une séquence pour une classe donnée.

Dans le processus de classification des miARNs, l'utilisation des attributs basés sur des caractéristiques de biogenèse ainsi des niveaux d'expression et des fonctions et non seulement sur des caractéristiques des positions sur les précurseurs diminuerait le TFP de la classification. Le TFP dépend aussi de la qualité de la classe négative. Plusieurs méthodes d'identification des miARNs et de leurs précurseurs

souffrent d'un TFP élevé (y compris MiRdup*). La classification des miARNs est un problème de classification binaire où la classe positive est représentée par des vrais miARNs et la classe négative par des faux miARNs. Plusieurs techniques peuvent être utilisées pour générer l'ensemble de données de la classe négative incluant 1) l'extraction aléatoire des sous-séquences de régions génomiques pauvres en miARNs (Kadri *et al.*, 2009), 2) l'extraction des sous-séquences de régions 3' non traduites des ARNm (Yousef *et al.*, 2006), 3) l'extraction des sous-séquences des régions codantes des gènes (Kadri *et al.*, 2009; Jiang *et al.*, 2007) ou 4) la permutation des positions des miARNs sur leurs précurseurs (Leclercq *et al.*, 2013). Ces techniques génèrent la classe négative dans l'espace de données d'entrée. La génération de la classe négative dans l'espace des attributs est une autre approche qui pourrait être utilisée dans la classification des séquences nucléotidiques.

APPENDICE A

DONNÉES SUPPLÉMENTAIRES DU CHAPITRE 2

Table A.1: Learning algorithms.

Algorithm type	Algorithm	Weka module	Options	Acronym
Symbolic	C4.5 decision tree	weka.classifiers.trees.J48	-C 0.25 -M 2	J48
	Random forests	weka.classifiers.trees.RandomForest	-I 10 -K 0 -S 1 -num-slots 1	RFT
Statistical	Naive bayes	weka.classifiers.bayes.NaiveBayes	NaiveBayes	NBA
	Support vector machine	weka.classifiers.functions.LibSVM	-S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model /home/hpvs/weka-3-7-10 -seed 1	SVM
	K-nearest neighbours	weka.classifiers.lazy.IBk	-K $K -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last""	IBK
Ensemble	AdaBoost	weka.classifiers.meta.AdaBoostM1	-P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 - -C 0.25 -M 2	ADA
	Bagging	weka.classifiers.meta.Bagging	-P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 - -C 0.25 -M 2	BAG

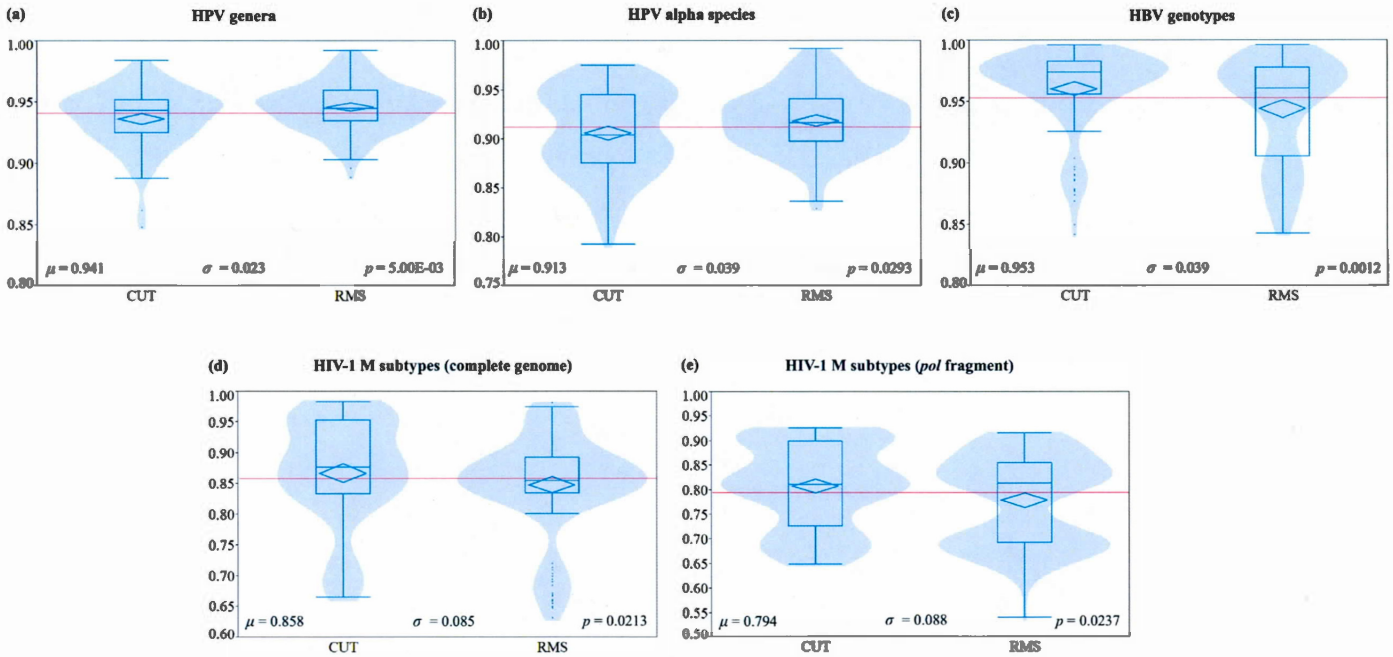


Figure A.1: Comparison of the weighted *F-measure* distribution according to *CUT* and *RMS* computed from the simulation study of the 280 experiments for **(a)** HPV genera, **(b)** HPV alpha species, **(c)** HBV genotypes, **(d)** HIV-1 M subtype complete genomes and **(e)** HIV-1 M subtype *pol* fragments. μ, σ are the mean and the standard deviation of the overall weighted *F-measures*. p is the *p-value* determining the statistically significance of the weighted *F-measure* mean differences among all the experiments. This *p-value* is computed with the Wilcoxon/Kruskal-Wallis test

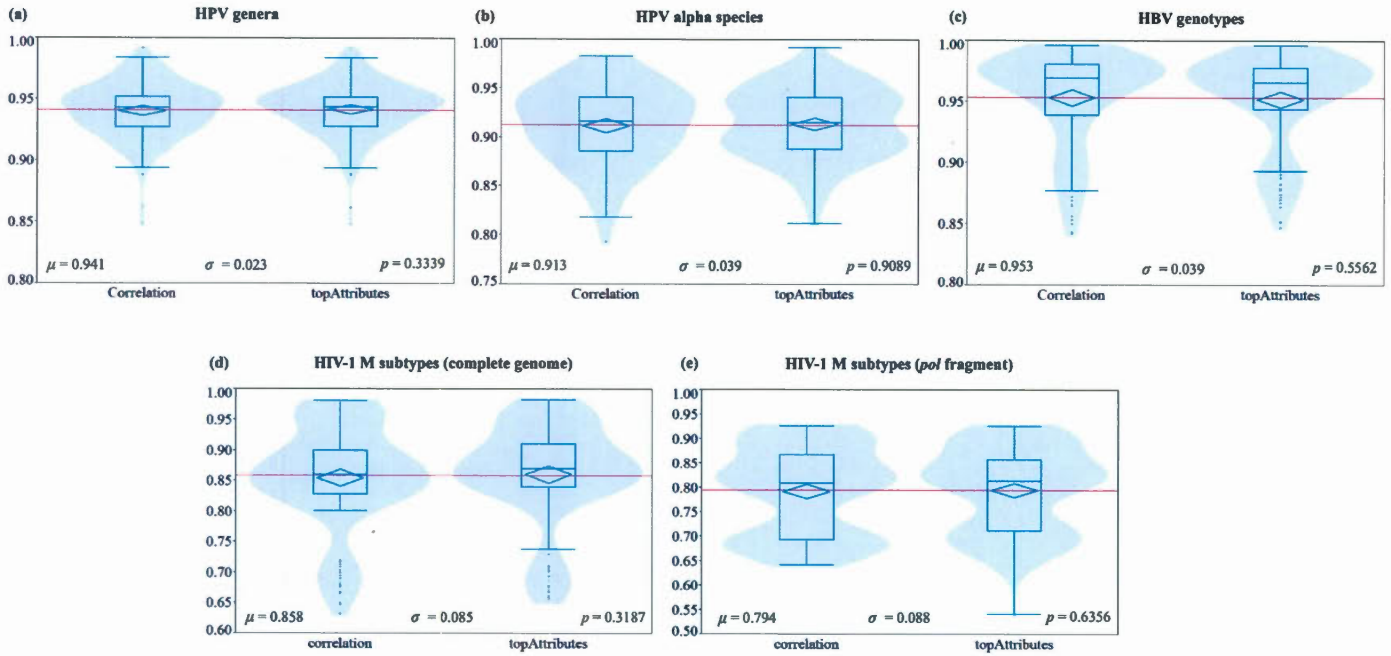


Figure A.2: Comparison of the weighted *F-measure* distribution according to *topAttributs* and *correlation* computed from the simulation study of the 280 experiments for **(a)** HPV genera, **(b)** HPV alpha species, **(c)** HBV genotypes, **(d)** HIV-1 M subtype complete genomes and **(e)** HIV-1 M subtype *pol* fragments. μ, σ are the mean and the standard deviation of the overall weighted *F-measures*. p is the *p-value* determining the statistically significance of the weighted *F-measure* mean differences among all the experiments. This *p-value* is computed with the Wilcoxon/Kruskal-Wallis test

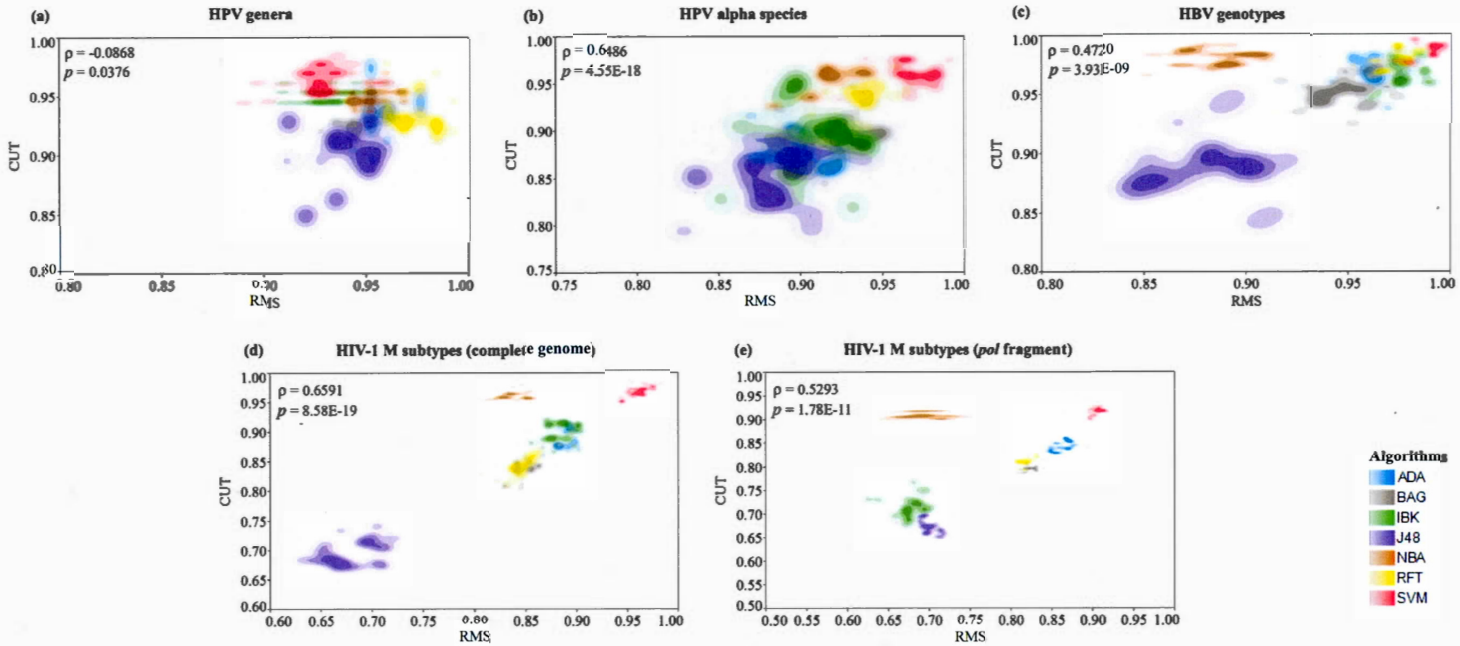


Figure A.3: CUT/RMS weighted *F-measure* correlation computed from the simulation study of the 280 experiments for **(a)** HPV genera, **(b)** HPV alpha species, **(c)** HBV genotypes, **(d)** HIV-1 M subtype complete genomes and **(e)** HIV-1 M subtype *pol* fragments. The weighted *F-measure* values are clustered according to the seven machine learning algorithms indicated in the legend. The values in top left corner of each figures correspond to the Spearman correlation score and the associated *p-value*

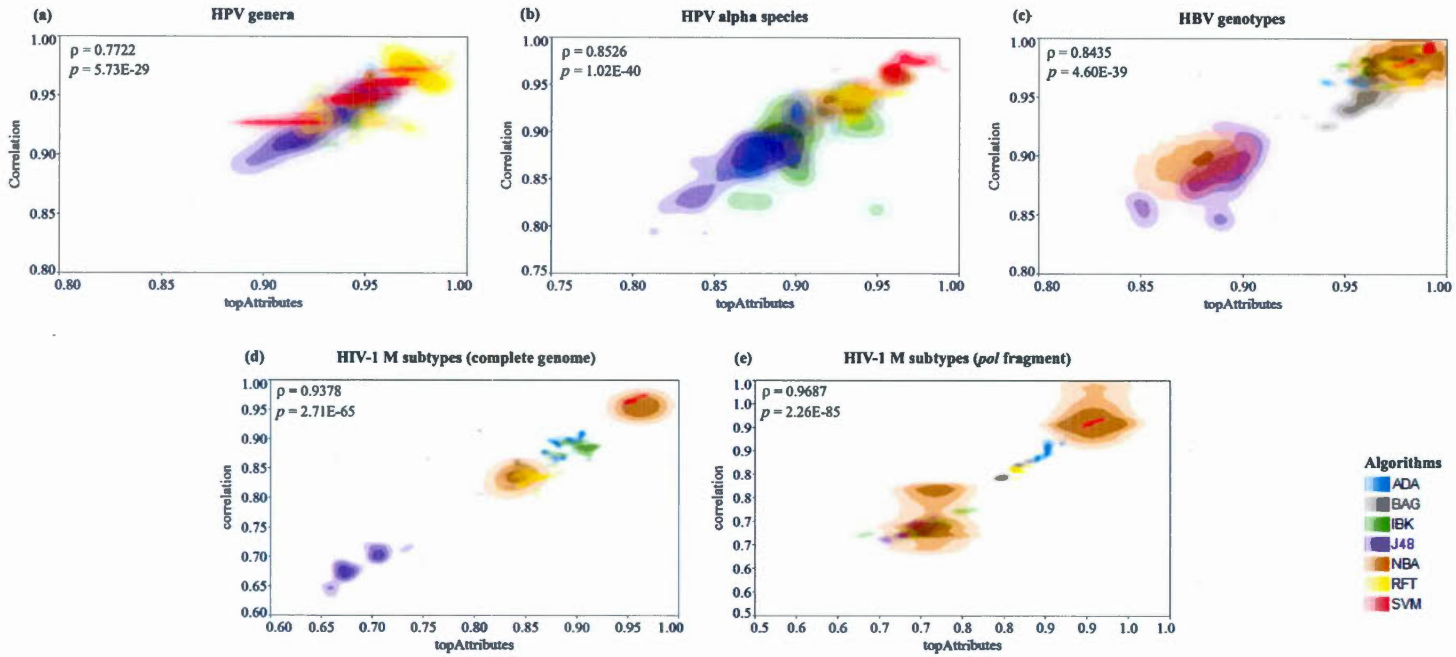


Figure A.4: *Correlation/topAttributes* Spearman correlations based on the weighted *F-measures* computed from the simulation study of the 280 experiments for **(a)** HPV genera, **(b)** HPV alpha species, **(c)** HBV genotypes, **(d)** HIV-1 M subtype complete genomes and **(e)** HIV-1 M subtype *pol* fragments. The weighted *F-measure* values are clustered according to the seven machine learning algorithms indicated in the legend. The values in top left corner of each figures correspond to the Spearman correlation score and the associated *p-value*

APPENDICE B

DONNÉES SUPPLÉMENTAIRES DU CHAPITRE 3

B.1 Supplementary Methods

B.1.1 Plant treatment

To identify miRNAs associated with stress responses and tolerance, a control library from plants grown under normal conditions was constructed except for the Al treated library from spring wheat Bounty (sensitive genotype). We used the control library from winter wheat Atlas66 (the tolerant genotype) as a control since in our previous study; the expression level of most genes was similar in control plants from Atlas66 and Bounty. Only 1.26% of genes were differentially expressed in the microarray analysis between Atlas66 and Bounty control demonstrating that most genes are expressed at a similar level (Houde et Diallo, 2008). For cold and salt treatments, seeds were grown in 3 L pots containing mixed soil (compost, black earth, vermiculite, 1 : 1 : 1) at 20°C under a 16 h photoperiod with a light intensity of 250 μmol m^{-2} s^{-1} and 70% relative humidity. To control the soil water content, one week before starting treatment, all plants were watered with the same amount of water, so they were exposed to the same soil moisture content. Control untreated plants were kept under the same conditions (normal temperature and watering) for the duration of the experiment, while treated plants were subjected to different abiotic stresses for various periods of time. For salt treatment, four

week-old plants were watered daily with the same volume of a 200 mM NaCl solution and sampled at days 1, 3, 5, 7, 12 and 15. For cold acclimation, two week old plants were transferred to a 4oC growth chamber in the same conditions of light and watering, and sampled after 1, 2, 7, 14, 21, 28, 35, 49 and 56 days. For Al treatment, seeds were grown in moist vermiculite during 5 d at 20°C under continuous light : 25 μmol m^{-2} s^{-1}. Seedlings were exposed to Al stress as described in Hamel *et al.* (1998) using 5 μM Al for Bounty and 50 μM Al for Atlas66.

B.1.2 Reference genome

We collected the Expressed Sequenced Tags (ESTs) from different databases : GrainGenes (Carollo *et al.*, 2005), TIGR (Childs *et al.*, 2007), WheatDB (wheatdb.ucdavis.edu :8080/wheatdb/), TAGI (compbio.dfci.harvard.edu/tgi/), Komugi (www.shigen.nig.ac.jp/wheat/komugi/) and from our database (Houde *et al.*, 2006). These data were clustered according to Uniprot UniRef100 clusters to reduce the noise given by partially and nearly identical ESTs. To cluster the ESTs, each one has been aligned with the program blastx (Gish *et al.*, 1993) with default parameters. Each EST is annotated with a Uniref100 protein if it has at least 60% identity, 60% query coverage and/or hit coverage and E-value inferior to 5E-5.

B.1.3 Removing adaptor and read mapping

All sequences are from SOLID sequencing and had a length of 35 with T as first character. Since expected mature miRNA length is around 21nt, reads should contain a part of the adaptor P2. Adapters were removed in color space (represented by numbers between 0 and 3) using the program cutadapt v0.9 (Schulte *et al.*, 2010) with the following parameters : -c -e 0.12 -a 330201030313112312 –maq.

Those parameters accept at most 12% of mismatches over the adaptor, therefore adjusting the number of mismatch depending on sequence length. The output format was adapted to the program MAQ (Ondov *et al.*, 2008) for mapping the adaptor free reads to the ESTs. Identical or different reads with the same mapping results (same region in the EST) have then been grouped as a small RNA. We computed the abundance of a small RNA as the sum of all reads that exhibit the given small RNA after mapping.

B.1.4 MiRdup* model creation

B.1.4.1 Collected data

All miRNAs and pre-miRNAs were downloaded from miRBase release 18, which contains 21,643 miRNAs or miRNAs* for 18,226 pre-miRNAs. The negative dataset was generated within the same data, except the start position of the miRNA was chosen randomly on the pre-miRNA. The different lengths were preserved. The negative miRNA cannot be equal to the original miRNA and cannot exist in the positive training dataset. The complete training dataset has 46,412 instances, so 23,206 instances for each positive and negative dataset. This equal representation avoids any influence by one or the other datasets on the model. In order to test our classifiers, we validated them on miRBase and a few other datasets. The first dataset, Plant miRNAs Database (Zhang *et al.*, 2010), which contains 10,081 instances of combination miRNA and pre-miRNA. Two other datasets come from papers which can analyze real data using different ways to predict miRNAs : MIRcheck (Jones-Rhoades et Bartel, 2004) with 187 instances and miRDeep (Friedländer *et al.*, 2008) with 228 instances.

B.1.4.2 Model creation

To perform the feature selection and create models, we used Weka and its libraries (Hall *et al.*, 2009). For all trained models, we applied a 10 fold cross validation. In MiRdup*, we chose the combination of Adaboost M1 method (Freund et Schapire, 1997) and forest of random trees (Breiman, 2001). Adaboost is a machine learning algorithm specialized in problem optimisation and the search for the best global optimum (Osman et Kelly, 1996). It is used in combination with many other machine learning algorithms in order to improve their performances. We trained it with 10 iterations, reweighting, and a weight threshold of 100. Concerning other classifiers, the SVM classifier, working with libSVM library (Chang et Lin, 2011), was trained with radial kernel and the C4.5 tree (Quinlan, 1993) was trained with Adaboost. Sensitivity and specificity, which measures the proportion of positives and negative instances which are correctly classified as such respectively, were calculated with the well-known formulas :

We set 35 features that can characterize the duplex miRNA and its complement sequence on the hairpin (Table B.5). To calculate them, our model requires the mature sequence of the miRNA, the pre-miRNA hairpin loop sequence and its secondary structure. Few are based on the position and length of particular structural elements in the miRNA, like bulges and bases pairs, or the situation of the miRNA itself in the hairpin, like the distance relative to the loop. We also included features that characterize the nucleotide sequence, where we check for missing nucleotides in the miRNA, percentage of presence of certain nucleotides and GC content. Furthermore, we calculate the minimum folding energy of the duplex. Those features reveal some configuration errors found in pre-miRNA prediction results. In Figure SM1b, we show examples of misplaced miRNAs (red) on a predicted pre-miRNA (black), where the miRNA sequence might extend in the loop

or in a big bulge. We assume that could affect the biological process and thus, conclude that the miRNA cannot be handled by the given pre-miRNA. Those should be invalidated. On the opposite, miRNAs (yellow) which have a good position on the hairpin are validated, depending of the training set characteristics.

B.1.4.3 Features ranking

Not all features have the same impact on classifiers. Some are too similar between the positive and negative datasets and can influence badly the classifier, especially in overfitting and computational time (Zhou *et al.*, 2006). Then, in order to get the most important features, we ranked them, as shown in Table B.5 for all the miRBase dataset. We observe that the most influent feature concerns the distance from the terminal loop and the weakest is the presence of U. In addition, presence and percentage in the miRNA of a certain nucleotide tend to be less influential features. The miRNA length is a positive control and has no difference between positive and negative datasets, because generation of the negative dataset is done by moving the miRNA while preserving its length. As shown by Leclercq *et al.* (2013), depending on the kingdom of species chosen for the study, the rank of important features will change. Hence, to ensure quality, we trained the model on all miRBase, all plants and monocot datasets.

B.1.4.4 Training models on classifiers

Various classifiers were tested in order to find the best one which could discriminate all experimentally validated miRNAs from miRBase dataset depending on the class (Table B.1) : SVM trained with radial basis kernel, boosting with Adaboost on C4.5 tree and Adaboost on forest of random trees. Two sets of features were experienced on the classifiers : best features and all features. Best overall scores are given by Adaboost on forest of random trees on all 35 features, with

80.23% correctly classified instances and 86.5% sensitivity (SE). Only specificity (SP) with 74.1% is less than with others classifiers. SVM showed 93.5% SP, but this classifier could only classify instances correctly as 71.81% and has a poor SE of 50.2%. Concerning boosting on C4.5 tree, results are close to forest of random trees, but are always slightly below. However, we noticed that SVM can get better scores when using only the best features set, and this improves correctly classified instances to 75.77% and SE at 71.1%. However, this is still less than forest on random trees with all features. It's important to note that only SVM improves its results with only best features, which is the opposite with other tested classifiers. Also, a disadvantage of the SVM is that it takes a very long time to train on all miRBase data, several hours compared to few minutes for other classifiers. Hence, the best compromise would be Adaboost on Random Forest.

B.1.4.5 Evaluation of the designed models

After training the classifiers on all features, few datasets were submitted to construct the models. These include miRBase and plant miRNAs (PMRD) databases, and published results from MIRcheck and miRDeep. Results are shown in Table B.2, where we have the total instances of every dataset, the portion of validated instances and the mean absolute error. An instance is a miRNA and its corresponding precursor. The mean absolute error informs about the average error per predicted instance made by the model. We note in the Table B.2 that the mean absolute error follows the percentage of validated instances. We also observe that Adaboost on forest of random trees model have the best scores of validated instances and the lowest mean error rate in every dataset, except MIRcheck, where Adaboost on C4.5 model validated 4 more instances. The most interesting is miRBase dataset, which is our reference, where Adaboost on forest of random trees validated 99.75% of the 23,206 miRNAs with their precursors, with a very high confidence score (lowest mean absolute error). PMRD, which contains a lot of computational

predictions, has only 78.18% of validated instances and makes models less confident with their predictions than the miRBase dataset. On the other hand, we can assume that MIRcheck already does very good predictions, with 94.65% of validated instances with Adaboost on C4.5 model, which is not the case of miRDeep, the best score of validated instances is 84.51%, predicted with Adaboost on forest of random trees model. Finally, SVM has lower scores than other models. Since we used the model trained on all features, which does not advantage SVM, results are improved if the model is trained on best features, but not enough to beat other classifiers on the miRBase dataset : 97.97% of correctly classified instances (22,736 instances) and a mean absolute error of 0.0203.

B.1.4.6 Conclusion of the MiRdup* model

We trained several classifiers, and boosting on forest of random trees showed best overall results. It has the best classification score of 80.29% and ability to identify negatives results (specificity) 86.5%. Its sensitivity is not the top, but still gives a high score of 74.1%. The evaluation of miRBase on this model has shown a high score of 99.75% validated instances of mature miRNAs and corresponding pre-miRNA(s). It's not the first time where SVM, Adaboost, trees and random forest are used in this field to classify real or false miRNAs and pre-miRNAs (Guan *et al.*, 2011; Helvik *et al.*, 2007; Jiang *et al.*, 2007; Zhou *et al.*, 2007). Adaboost on tree kept a high score level in this study, but is always lower than forest of random trees. Compared to these classifiers, the SVM has the limit of the training time, where several hours are needed to train the model. This is a problem to try the classifier with diverse parameters, and would be a disadvantage to the usage of MiRdup*. Perhaps several tests on parameters could improve its overall results. We know that relying only on miRBase could limit the discovery of novel miRNAs. However, we can assume that this database contains many examples of what characterizes a miRNA on its pre-miRNA. Machine learning algorithms

are very useful for this kind of problem. They can accept certain diversity around usual characteristics in order to assess new members that fulfill the majority of the statistics. Also, the majority of hairpins in miRBase are single loop hairpins, but several hundred contain multiloops. We designed this algorithm so that we can accept multiloop hairpins, since we only need to focus on the duplex containing the miRNA and its complement. Consequently, if plants contain more miRNAs with multiloop pre-MiRNAs, it won't affect MiRdup*. We noticed in this study that general statistics on features change between set of species, especially between different groups of plants. Those differences denote that specific datasets should be used depending on the hairpin predictions we want to validate. The objective of MiRdup* here is to reduce a noise as post treatment of predicted hairpins.

B.2 Supplementary Data

Supplementary Data files are downloadable from https ://static-content.springer .com/esm/art%3A10.1186%2Fs12864-015-1490-8/MediaObjects/12864_2015 _1490_MOESM4_ESM.zip

B.2.1 Data SD1

The file **SD1_viewMicroRNAinPremicroRNA.txt** exhibits the miRNA in the context of their folded pre-miRNA in association with all the small RNAs expressed in the given context. The predictors (HHMMiR or MiPred), the dataset training used by MiRdup* for the prediction of miRNA position in the precursor and the obtained score for each dataset are presented for each precursor. Line starts with (#) : Precursor sequence. Lines start with (&) : Secondary structure of the precursor(s) in dot-bracket notation; Precursor id (apPre_xxxx); accesssion number of EST(s) producing the precursor(s), Uniref ID if available. Lines start with (>) : MiRNA sequence; Expression level of the miRNA in the 10 libraries

corresponding to the number of reads sequenced in library L1 to L10 ; [total expression of miRNA candidates] ; apMir xxxx, miRNA apMir ID ; md, MiRdup* ; mdB, mdP and mdM represents dataset training used by MiRdup* to predict the position of miRNA candidate in the precursor : mdB, md trained on miRBase (B) ; mdP, md trained on plants (P) ; mdM, md trained on monocots (M) ; mdB, mdP or mdM _Precursor ID_MiRdup score, the dataset training_ the Id of the precursor(s) from which the miRNA is predicted_the prediction score of MiRdup ; HHMMiR, and/or MiPred, the predictors used to predict the secondary structure ; +/-, the strand. Lines start with (%) : small RNA (sRNA) sequence mapping the precursor ; Expression level of the sRNA in the 10 libraries corresponding to the number of reads sequenced in library L1 to L10 ; [total expression of small RNAs] ; MiRNA apMir ID if this sRNA is predicted by HHMMiR or MiPred ; Precursor apPre ID if this sRNA is predicted with the current precursor with HHMMiR or MiPred ; +/-, strand. consmRNA_xxxx miRxxx, apMir having sequence homology (0-2 mismatches) with a given miRNA family from miRase(v21).

B.2.2 Data SD2

The file **SD2_MicroRNAsAbundancesSmall.txt** exhibits sorted miRNAs according to their percentage of overall abundance of small RNAs covering the hairpins. They are divided in three groups : higher than 50% ([100..51]), between 50% and 30% [50..31], below 30% [30..10]. #, pre-miRNA ID ; >, miRNA sequence mapped in a given pre-miRNA ; %, small RNA sequence mapped in a given pre-miRNA ; ( ), represents the number of reads in each sequenced library ; , represents the total of reads in the 10 sequenced libraries. For detailed information about the libraries and conditions see Method B.1.1 and Table B.3.

B.2.3 Data SD3

The file **SD3_MicroRNAstotalAbundancesInAllSequencedLibraries.txt** exhibits the miRNAs classified based on their total abundance in the ten libraries.

B.3 Gene ontology enrichment for predicted target genes

GO enrichment analysis of predicted miRNAs targets in separate and all combined libraries revealed that they may localize in diverse cellular compartments, play various functions in diverse biological and physiological processes. Targets related to the cell component category are associated with 21 GO Slim terms from which the term nucleus (P-value = 9.1e-004) shows enrichment in the development library (L3) suggesting their possible implication in regulating gene expression during reproductive phase. Target genes related to the molecular function category are represented by 19 GO Slim terms from which six show significant enrichment (Table B.9 and Figure B.5a). The most enriched terms for all the libraries is lipid binding activity (P-value = 7.6e-006) and protein binding activity (P-value = 1.2e- 006). They include specific regulatory proteins and cell metabolism enzymes. Hence, the miRNA candidates are mainly associated with the modulation of several transcription factors, histones proteins and cellular enzymes involved in lipid metabolism and oxidative stress. For the targets from the roots libraries (L5 and L8 to L10), a significant enrichment is found for protein binding activity (P-value ranger from 2.4e-004 to 5.3e-007) and DNA binding activity (P-value ranger from 1.0e-003 to 1.0e-005) suggesting their possible implication in regulation of gene expression during roots development. Targets related to the biological process category are classified into 39 GO Slim terms, out of which six show enrichments. Targets from all the ten libraries are enriched for secondary metabolic process (P-value = 3.7e-005) whose related targets are known to function in both normal and stress conditions. The libraries L1 and L3 to L10 exhibited significant

enrichment for response to endogenous stimulus overrepresented by auxin responsive proteins (P-value ranger from 7.8e-002 to 4.4e-007) (Table B.9 and Figure B.6a). Consistent with the investigated conditions, the targets from vernalized library (L2) corresponding to the floral transition shows a significant enrichment for flower development (P- value = 2.9e-005) and those from the reproductive library (L3) are enriched for multicellular organismal development (P-value = 1.7e-005). These include several genes with various functions in stress responses and meristem initiation, regulation of flower development, root morphogenesis and leaf development (Table B.9 and Figure B.6a). These results support the possible function of the identified miRNAs in stress response and plant development.

B.4 Supplementary Tables

Table B.1: Results of various classifiers on the all miRBase training dataset with 10 fold cross validation

Classifier	Correctly classified instances	Correctly classified instances (%)	True positive rate	False positive rate	True negative rate	False negative rate	Sensitivity	Specificity
a)								
SVM radial basis kernel	33,332	71.81	0.502	0.065	0.935	0.498	0.502	0.935
Adaboost on C4.5 tree	36,780	79.24	0.814	0.229	0.771	0.186	0.814	0.771
Adaboost on Random Forest	37,262	80.28	0.865	0.259	0.741	0.135	0.865	0.741
b)								
SVM radial basis kernel	35,170	75.77	0.711	0.196	0.804	0.289	0.711	0.804
Adaboost on C4.5 tree	36,157	77.9	0.802	0.244	0.756	0.198	0.802	0.756
Adaboost on Random Forest	36,646	78.958	0.853	0.274	0.726	0.147	0.853	0.726

Best scores for a classifier in each category are highlighted in bold. **a**. Classifiers trained on all 35 features. **b**. Classifiers trained on the 14 best features.

Table B.2: Evaluation of chosen models trained on all features

Dataset	miRBase		PMRD		MIRcheck		miRDeep	
Total instances	23,206		10,081		187		226	
Model	Validated instances	Mean absolute error	Validated instances	Mean absolute error	Validated instances	Mean absolute error	Validated instances	Mean absolute error
SVM radial basis	96.51%	0.03	36.91%	0.63	59.35%	0.4	52.21%	0.47
Adaboost on C4.5 tree	99.49%	0.005	72.32%	0.28	94.65%	0.06	80.53%	0.2
Adaboost on Random Forest	99.75%	0.002	78.18%	0.21	92.51%	0.07	84.51%	0.16

The evaluation is done on predicted hairpins composed of predicted pre-miRNAs with their sequenced miRNAs. Data comes from miRBase, PMRD, MIRcheck and miRDeep.

Table B.3: Description of the 10 libraries (L1 to L10)

	Genotype		Phases		Tissues			Abiotic stress				Growth conditions and sampled time points
	T	S	V	R	A	Rt	Sp	CV	Sa	Al	Nc	
L1	×		×		×						×	Normal conditions (Nc) : one, two, and three week-old plants grown at 20°C with normal watering
L2	×		×		×			×				Vernalization (V) : two week-old plants acclimated at 4°C for 1, 2, 7, 14, 21, 28, 35, 49, 56 days
L3	×			×	×		×				×	Normal conditions : 56 day vernalized plants transferred to 20°C for 3-6 weeks
L4			×		×				×			Salt (Sa) : three week-old plants of the genotype Clair watered with 200 mM of NaCl solution for 1, 3, 5, 7, 12 and 15 days
L5			×			×			×			
L6		×	×		×						×	Normal conditions (Nc) : 1, 2, and 3 week old plants grown at 20°C with normal watering
L7		×	×		×			×				Cold (C) : two week old plants acclimated at 4°C for 7, 14, 21, 28 days
L8		×	×			×				×		Al (Al) : one week old seedlings exposed to 5 μM Al for 1 day
L9	×		×			×					×	Normal conditions (Nc) : root tissue of 1 week-old seedlings with normal watering
L10	×		×			×				×		Al (Al) : one week-old seedlings exposed to 50 μM Al for 1 day

The library number corresponds to the number of the barcode used in the cDNA construction. Different tissues including aerial parts (A) which comprise leaves and crowns, spikes (Sp) and roots (Rt) from tolerant (T) or sensitive (S) genotypes in vegetative (V) or reproductive (R) phase under normal conditions (Nc) or subjected to short exposure to cold (C), long exposure to cold (vernalization ; V), salt (Sa) or Al (Al). The tolerant genotypes are Clair and Atlas for cold and Al, respectively and the sensitive genotype is Bounty for both stresses. For more details, see Method B.1.1.

Table B.4: Quality values (QV) of predicted miRNAs color reads based on analyses of the quality files provided by SOLID sequencing in the first 10 color bases

QV<10	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	Mean
0	95,73	94.74	96.03	94.20	94.72	94.46	93.36	95.22	95.78	94.11	94.74
1	0.57	0.63	0.47	0.74	0.52	0.64	0.73	0.61	0.70	0.80	0.64
2	0.52	0.62	0.40	0.70	0.60	0.60	0.72	0.63	0.58	0.71	0.61
3	0.44	0.54	0.34	0.60	0.46	0.55	0.67	0.47	0.42	0.63	0.53
4	0.46	0.52	0.39	0.70	0.53	0.54	0.62	0.51	0.44	0.58	0.54
5	0.39	0.51	0.35	0.59	0.49	0.54	0.66	0.50	0.39	0.55	0.51
6	0.38	0.50	0.41	0.50	0.46	0.53	0.64	0.47	0.38	0.49	0.49
7	0.37	0.46	0.41	0.53	0.49	0.51	0.63	0.46	0.33	0.51	0.48
8	0.38	0.48	0.40	0.49	0.56	0.54	0.66	0.42	0.34	0.54	0.49
9	0.38	0.48	0.35	0.45	0.54	0.54	0.65	0.35	0.32	0.50	0.47
10	0.37	0.52	0.45	0.50	0.63	0.54	0.67	0.35	0.33	0.58	0.50

QV<10 is the number of bases having a quality value (QV) strictly lower than 10 in the first 10 bases. First column contains the number of possible bases with quality value strictly inferior to 10 that is considered as high sequence quality. Results are shown for the ten sequenced libraries. The ten columns (L1-L10) contain the percentage of reads sequences in each library having quality value in bases inferior to 10; the last column corresponds to the mean in all libraries. For detailed information about the libraries and conditions see Method B.1.1 and Table B.3.

Table B.5: The 35 miRNA features used by MiRdup* to classify pre-miRNA candidates and their importance in the whole prediction

Rank score	ID	Features
0.27457256	22	Distance from terminal loop
0.22249313	25	Length of overlap in loop
0.19002714	28	Average number of paired bases in window 3
0.18625003	35	Length of biggest bulges in percentage of the miRNA length
0.18453715	34	Length of biggest bulge
0.17837146	27	Average number of paired bases in window 5
0.16974265	7	Number of base pairs in duplex
0.16260705	26	Average number of paired bases in window 7
0.15725944	2	Duplex minimum folding energy
0.11455883	24	miRNA included in loop
0.07013466	23	Distance from hairpin start
0.0648666	13	Start of perfect 5 mer base pair
0.05432927	5	Maximum length without bulges in percentage of the miRNA length
0.05422537	4	Maximum length without bulges
0.0262861	11	Start of perfect 10 mer base pair
0.02460451	6	Length without bulges from miRNA start
0.02011416	12	Presence of perfect 5 mer base pair
0.01970081	10	Presence of perfect 10 mer base pair
0.01560982	29	Bulge at position 2
0.01496887	3	GC percentage
0.01059411	30	Bulge at position minus 2
0.00826505	32	Bulge at position minus 1
0.00707627	20	Percentage of G in the miRNA
0.00453195	9	Start of perfect 20 mer base pair
0.00435343	8	Presence of perfect 20 mer base pair
0.00402892	33	Number of bulges
0.00363068	18	Percentage of A in the miRNA
0.0029013	19	Percentage of U in the miRNA
0.00164901	21	Percentage of C in the miRNA
0.001163	17	Presence of C
0.00042975	31	Bulge at position 1
0.00007034	16	Presence of G
0.00002714	14	Presence of A
0.00000199	15	Presence of U
0	1	miRNA length

Features ranking on all sequenced or cloned miRNAs from miRBase dataset with the ranker of InfoGain. For each instance containing the miRNA, pre-miRNA and the secondary structure, 35 features are extracted. IDs were set arbitrarily. Gray area represents the 14 best features having a rank superior to 0.05.

Table B.6: Quality values (QV) of predicted miRNAs color reads (corresponding to miRNA candidates) based on analyses of the quality files provided by SOLID sequencing in the first 10 color bases

Bases with QV<10	Number of reads	% of mapping	Predicted miRNAs	% of predicted miRNAs	Reads	% of reads
0	50031306	56.15	128	64.32	833611	94.82
1	13518710	15.17	4	2.01	5556	0.63
2	9281699	10.42	5	2.51	5323	0.61
3	5812834	6.52	5	2.51	4589	0.52
4	4189907	4.70	3	1.51	4694	0.53
5	2426248	2.72	4	2.01	4424	0.50
6	1641104	1.84	3	1.51	4216	0.48
7	915577	1.03	6	3.02	4133	0.47
8	862935	0.97	7	3.52	4206	0.48
9	273630	0.31	7	3.52	4076	0.46
10	151146	0.17	27	13.57	4343	0.49

QV<10 is the number of bases having a quality value (QV) strictly lower than 10 in the first 10 bases. Since one predicted miRNA can come from many reads, we calculate the average QV<10 and round it down to integer. First column contains the number of possible bases with quality value strictly inferior to 10 that is considered as high sequence quality. The second column is the number of reads of the deep sequencing. The third column is the corresponding percentage of mapping. The fourth column is the number of predicted miRNAs. The fifth column is the percentage of predicted miRNAs compared to 1369. The sixth column is the number of reads (several reads can have the same sequence) and the seventh column is the corresponding percentage. For detailed information about the libraries and conditions see Method B.1.1 and Table B.3.

Table B.7: The different explored thresholds (Evalue) and Query/Hit coverage and percentage identity of ESTs producing the identified pre-miRNAs

a)

Evalue $\leq$	percent identity $\geq$	HC $\geq$	Pre-miRNAs overlaping TE	miRNAs overlaping TE
0.1	80	0	136	29
0.00005	80	85	74	13
0.00005	95	85	31	13
1.00E-20	80	0	97	0
1.00E-20	80	85	68	0
1.00E-20	95	85	31	0
EV $\leq$ 0.00005 and EV $\geq$ 1E-20	80	0	116	13

b)

Evalue $\leq$	QC or HC $\geq$	percentage identity $\geq$	ESTs aligned with protein	ESTs producing miRNAs overlapping proteins	Number of miRNA overlapping proteins
	0	0	320	171	28
5.00E-05	0	0	308	163	27
5.00E-05	85	75	77	54	9
1.00E-20	0	0	232	148	20
1.00E-20	85	75	76	54	9
EV $\leq$ 0.00005 and EV $\geq$ 1E-20	0	0	221	77	22

ESTs are aligned (blasted) against a) TREP database for transposable elements; and b) plant proteins database. Results of blast are presented. For the remaining analysis, we retained 1.00E-20 and Query Coverage (QC) or Hit Coverage (HC) $\geq$ 85; and percentage of identity $\geq$ 75 for proteins.

Table B.8: List of predicted target genes and their associated Uniref and GO Slim terms when available. See excel file **TableS6_TargetGenes.xlsx** downloadable from https ://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-1490-8/MediaObjects/12864_2015_1490_MOESM2_ESM.zip.

Table B.9: Enrichment of GO Slim terms in the three gene ontology categories (cell component, molecular function and biological process) for targets of all miRNAs predicted from the ten sequenced libraries. See excel file **TableS7_GeneOntolgyEnrichmentAnalysis.xlsx** downloadable from https ://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-1490-8/MediaObjects/12864_2015_1490_MOESM2_ESM.zip.

Table B.10: The number of miRNA abundance level per library

L1				**L2**			
Level	0-30	31-50	51-100	Level	0-30	31-50	51-100
Low	9	7	54	Low	7	12	60
Medium	3	9	52	Medium	8	9	67
High	0	1	16	High	0	2	23
	12	17	122		15	23	150
L3				**L4**			
Level	0-30	31-50	51-100	Level	0-30	31-50	51-100
Low	5	9	63	Low	9	14	77
Medium	1	2	22	Medium	2	6	46
High	0	0	6	High	0	0	16
	6	11	91		11	20	139
L5				**L6**			
Level	0-30	31-50	51-100	Level	0-30	31-50	51-100
Low	7	10	78	Low	6	7	52
Medium	6	9	43	Medium	2	7	50
High	0	0	7	High	0	2	16
	13	19	128		8	16	118
L7				**L8**			
Level	0-30	31-50	51-100	Level	0-30	31-50	51-100
Low	9	8	66	Low	9	8	67
Medium	2	9	43	Medium	4	10	50
High	0	1	14	High	1	0	8
	11	18	123		14	18	125
L9				**L10**			
Level	0-30	31-50	51-100	Level	0-30	31-50	51-100
Low	7	12	73	Low	8	9	67
Medium	6	7	39	Medium	4	8	44
High	1	0	5	High	2	0	7
	14	19	117		14	17	118

(low, 10-99 reads; medium, 100-999 reads; and high, 1000 and more) and the percentage of the overall abundance of the given miRNA and others small RNAs mapped or positioned in a given pre-miRNA. The percentage of abundance of miRNA candidates is partitioned in three groups (low, medium, high) indicating the percentage between respectively [100..51], [50..31] and [30..0] of a given pre-miRNA. For detailed information about the libraries and conditions see Table B.3.

Table B.11: Number and characteristics of differentially expressed miRNAs under different growth conditions

	Conditions	Libraries	Number of expressed miRNAs	Number of DE miRNAs	Maximum FC up	Maximum FC down
Abiotic stress responses	Vernalization (winter wheat)	L2/L1	199	67	72	25
	Cold (spring wheat)	L7/L6	198	34	17	7
	Al (winter wheat)	L10/L9	192	85	23	39
	Al (spring wheat)	L8/L9	191	86	14	30
	Salt in leaves (winter wheat cv Clair)	L4/L1	198	55	82	41
	Total unique abiotic stress responsive miRNAs		-	165	-	-
Tolerance	Cold tolerance	L2/L7	199	52	17	21
	Al tolerance	L10/L08	190	27	4	23
	Total unique miRNAs		-	69	-	-
Development responses	Floral transition	L3/L2	199	76	18	67
	Flowering	L3/L1	194	56	63	17
	Total unique development responsive miRNAs	-	-	99	-	-
	Total unique		199	182	-	-

Normalized reads were compared between normal conditions and treatments (cold, salt and Al) for a given genotype to identify miRNAs associated with stress responses and between tolerant (winter wheat) and sensitive (spring wheat) genotypes to identify miRNAs associated with tolerance. For development, normalized reads were compared between plants in vegetative and reproductive phases. MiRNAs showing a fold change (FC) of two or more with an adjusted p-value < 0.05 in a given condition are presented. DE miRNAs are differentially expressed miRNAs

Table B.12: Grouping miRNAs based on their digital gene expression patterns

Investigated conditions and tissues types	MiRNA groups	Expression patterns	Number of miRNAs in each group
	Cold tolerance		
Cold/vernalization (Aerial parts)	Co1	not_L2/L1_up_L7/L6	11
	Co2	not_L2/L1_dw_L7/L6	12
	Co3	up_L2/L1_not_L7/L6	39
	Co4	dw_L2/L1_not_L7/L6	17
	Co7	up_L2/L1_dw_L7/L6	1
	Co8	dw_L2/L1_up_L7/L6	1
	Cold response		
	Co5	up_L2/L1_up_L7/L6	8
	Co6	dw_L2/L1_dw_L7/L6	1
Aluminum (root tips)	Al1	not_L10/L9_up_L8/L9	19
	Al2	not_L10/L9_dw_L8/L9	5
	Al3	up_L10/L9_not_L8/L9	14
	Al4	dw_L10/L9_not_L8/L9	9
	Al7	up_L10/L9_dw_L8/L9	0
	Al8	dw_L10/L9_up_L8/L9	2
	Al responses		
	Al5	up_L10/L9_up_L8/L9	42
	Al6	dw_L10/L9_dw_L8/L9	18
Development (vegetative and reproductive tissues)	Dev1	not_L3/L2_up_L3/L1	11
	Dev2	not_L3/L2_dw_L3/L1	12
	Dev3	up_L3/L2_not_L3/L1	5
	Dev4	dw_L3/L2_not_L3/L1	38
	Dev5	up_L3/L2_up_L3/L1	8
	Dev6	dw_L3/L2_dw_L3/L1	20
	Dev7	up_L3/L2_dw_L3/L1	1
	Dev8	dw_L3/L2_up_L3/L1	4

Table B.13: Oligonucleotides used as probes in northern blot analysis.

MiRNAs ID	Oligonucleotide Sequence (5'-3')	Length nt
MiRNAs predicted in common by MiRdup* and MIRcheck (CM*M)		
apMir_20602	CTCCGTTCCAATATAGATGAC	21
apMir_19980	GGGTGATGGATGATCGATG	19
apMir_14769	TCAATACATATATGACAAC	19
apMir_21052	GTATTGGGTAATCTCATCTCA	18
Predicted MiRNAs specific to MiRdup* (SM*)		
apMir_16808	TGGTAGGATGGATGATGCTAT	21
apMir_86202	ACGGGCCGCACCGCTGGCCGACCCT	25
apMir_54471	GTGCCGGATTATGACTGA	18
Conserved miRNAs predicted in common by MiRdup* and MIRcheck (CM*M)		
apMir_22246 (tae-miR160a)	TGGCATACAGGGAGCCAGGCA	21
apMir_20968 (miR395a-21)	AGAGTTCCCCCAAACACTTCA	21

B.5 Supplementary Figures

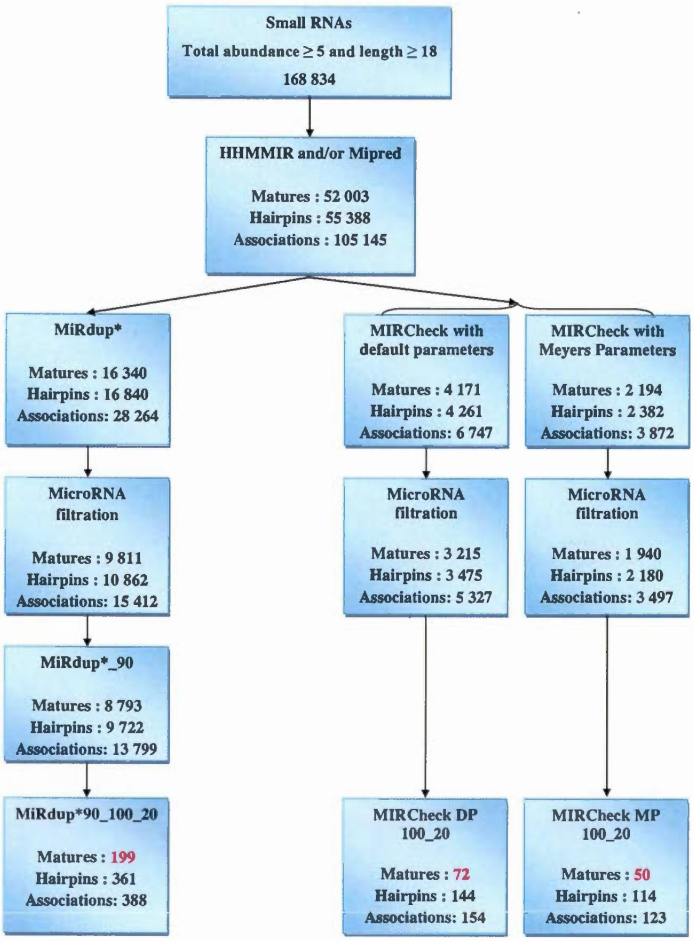


Figure B.1: The miRNA filtering pattern comparing MiRdup* trained on all experimental miRNAs of miRBase to MIRcheck default parameters (DP) and MIRcheck with rules from Meyers *et al.* (2008) (MP).

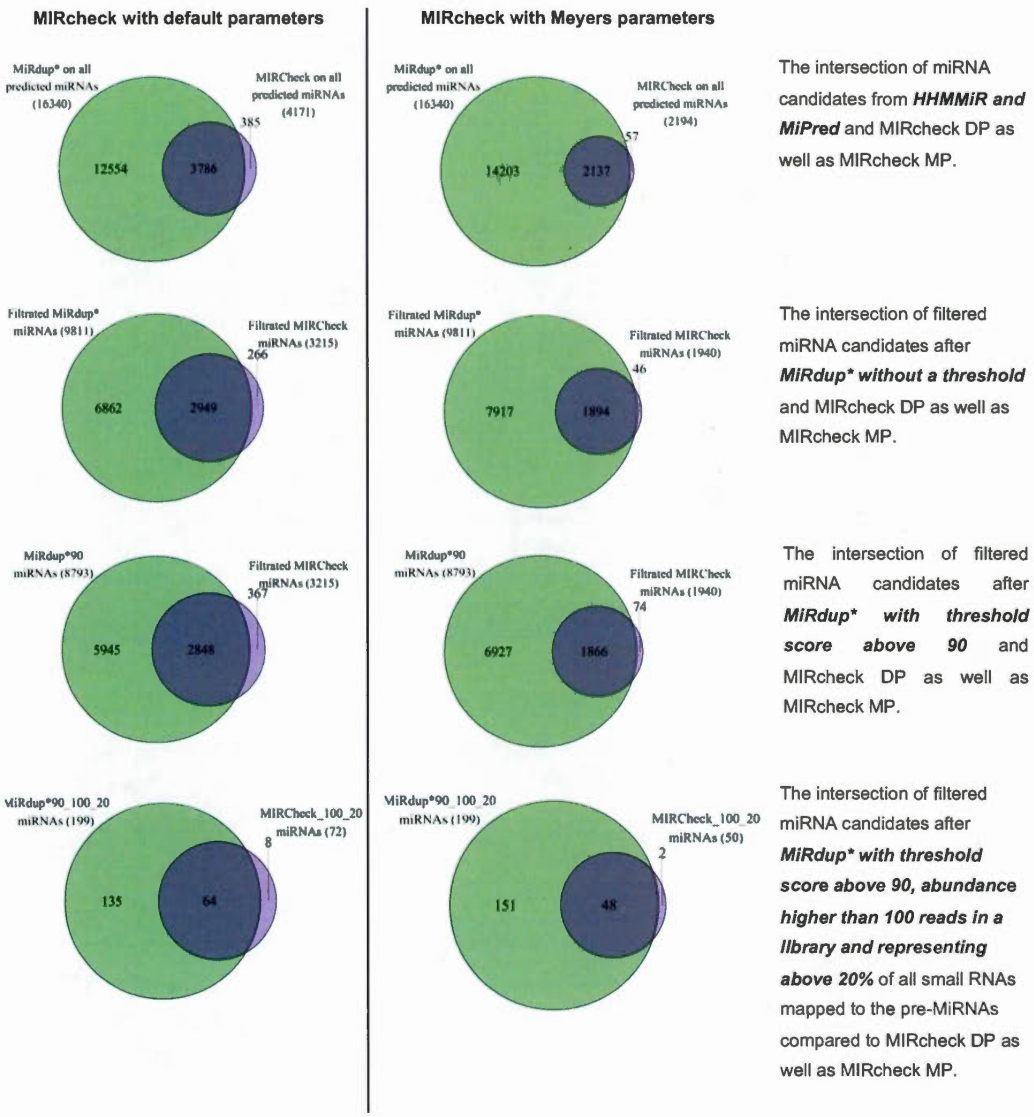


Figure B.2: The Predicted miRNAs intersection between the methods MiRdup*, MIRcheck with default parameters (DP) and MIRcheck with rules from Meyers *et al.* (2008) (MP).

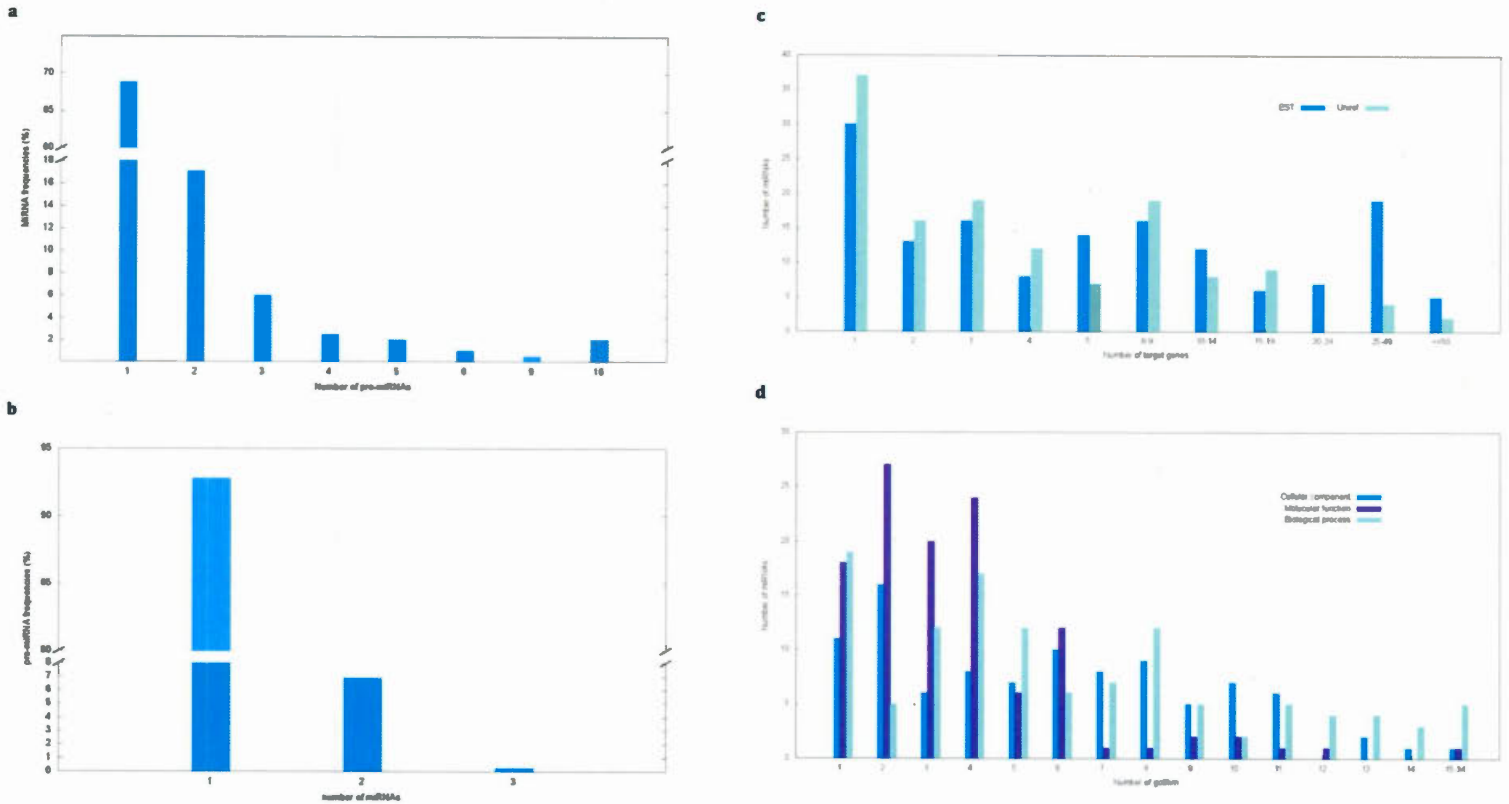


Figure B.3: Main characteristics of predicted miRNAs. a) Percentage of predicted miRNAs with a given number of pre-miRNA ; b) Percentage of pre-miRNAs carrying a given number of distinct predicted miRNAs ; c) Fraction of predicted miRNAs targeting a given number of ESTs or UniRef ids ; d) the fraction of predicted miRNAs targeting a given number of GO Slim term in the 3 main Gene Ontology categories. The number and the description of these GO Slim terms are presented in Figures B.4a-B.6a and Table B.9

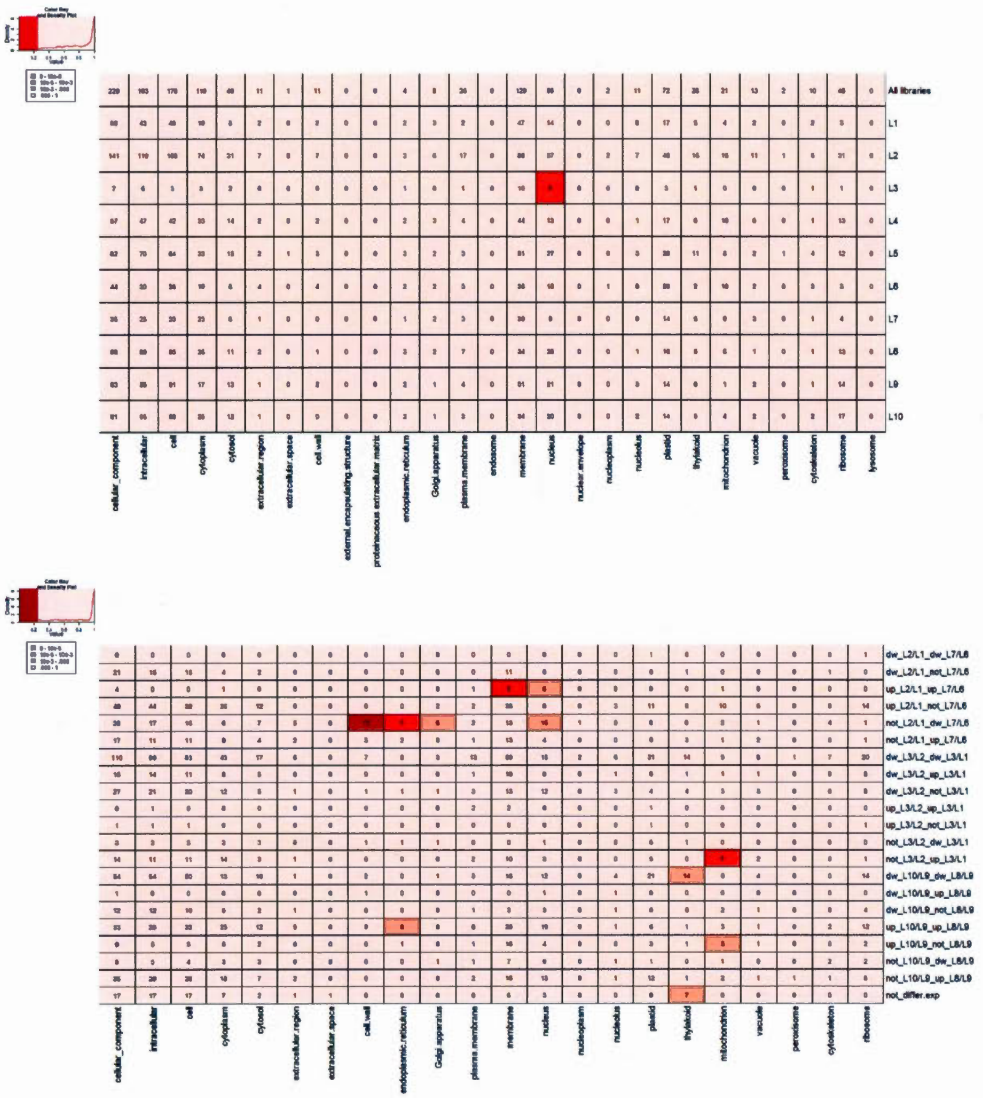


Figure B.4: Enrichment of main cell component GO Slims. a) Enrichment for the target genes in different libraries. b) Enrichment for the target genes of regulated miRNA candidates, under cold, Al, and development, grouped on 24 groups (eight per investigated condition) based on their differential expression patterns described in Table B.13. Dw,down regulated; up, up-regulated; not, not regulated. The value in each case indicates the number of miRNA-GO associations for the corresponding GO Slim.Targets associated with more than one cell component were assigned to more than one GO Slim term. The enrichment is presented in four different colors ("brown square symbol" high enrichment (P-value $< 10^{-5}$), "orange square symbol" medium enrichment ($P - value < 10^{-3}$), "light orange square symbol" low enrichment ($P - value < 0.05$) and "white square symbol" no enrichment ($P - value = 0.05$)). MiRNA groups having targets not annotated with a UniRef are not presented.

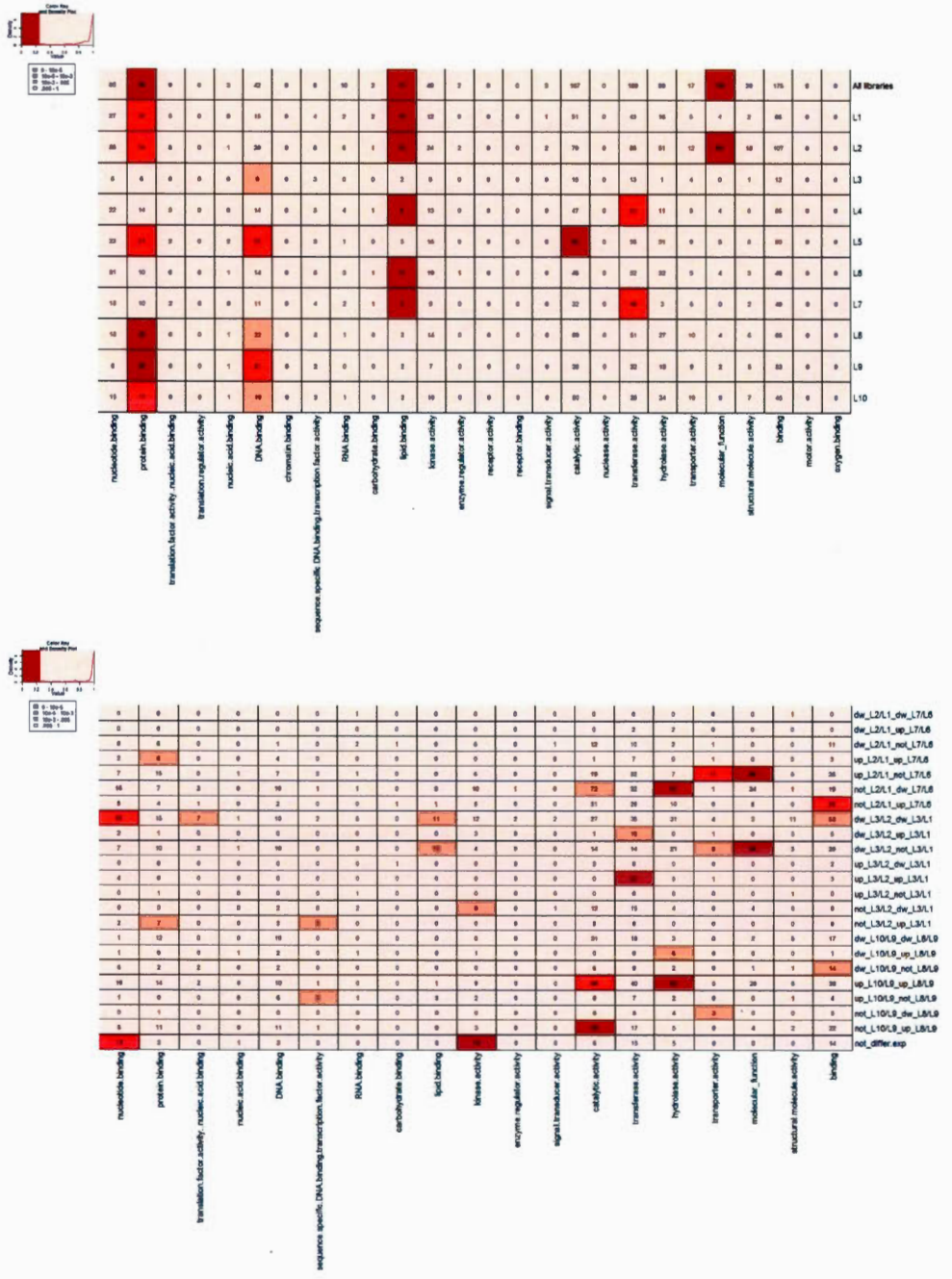


Figure B.5: Enrichment of main Molecular Function GO Slims.a) Enrichment for the target genes in different libraries. b) Enrichment for the target genes of regulated miRNA candidates, under cold, Al, and development, grouped on 24 groups (eight per investigated condition) based on their differential expression patterns described in Table B.13. Dw,down regulated ; up, up-regulated ; not, not regulated. The value in each case indicates the number of miRNA-GO associations for the corresponding GO Slim.Targets associated with more than one cell component were assigned to more than one GO Slim term.The enrichment is presented in four different colors ("brown square symbol" high enrichment (P-value $< 10^{-5}$), "orange square symbol" medium enrichment ($P - value < 10^{-3}$), "light orange square symbol" low enrichment ($P - value < 0.05$) and "white square symbol" no enrichment ($P - value = 0.05$)). MiRNA groups having targets not annotated with a UniRef are not presented.

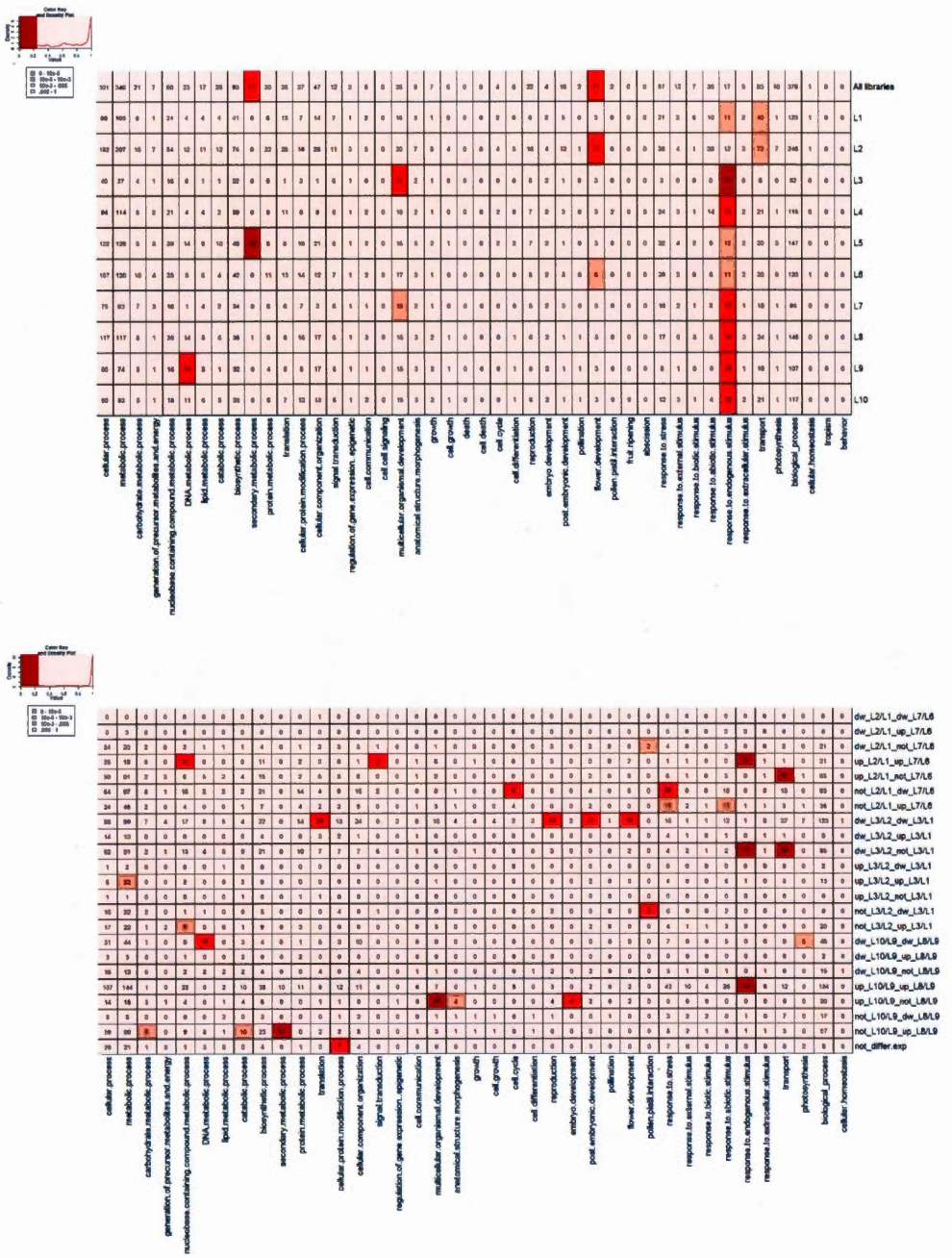


Figure B.6: Enrichment of main Biological Process GO Slims. a) Enrichment for the target genes in different libraries. b) Enrichment for the target genes of regulated miRNA candidates, under cold, Al, and development, grouped on 24 groups (eight per investigated condition) based on their differential expression patterns described in Table B.13. Dw,down regulated ; up, up-regulated ; not, not regulated. The value in each case indicates the number of miRNA-GO associations for the corresponding GO Slim.Targets associated with more than one cell component were assigned to more than one GO Slim term.The enrichment is presented in four different colors ("brown square symbol" high enrichment (P-value $< 10^{-5}$), "orange square symbol" medium enrichment ($P - value < 10^{-3}$), "light orange square symbol" low enrichment ($P - value < 0.05$) and "white square symbol" no enrichment ($P - value = 0.05$)). MiRNA groups having targets not annotated with a UniRef are not presented.

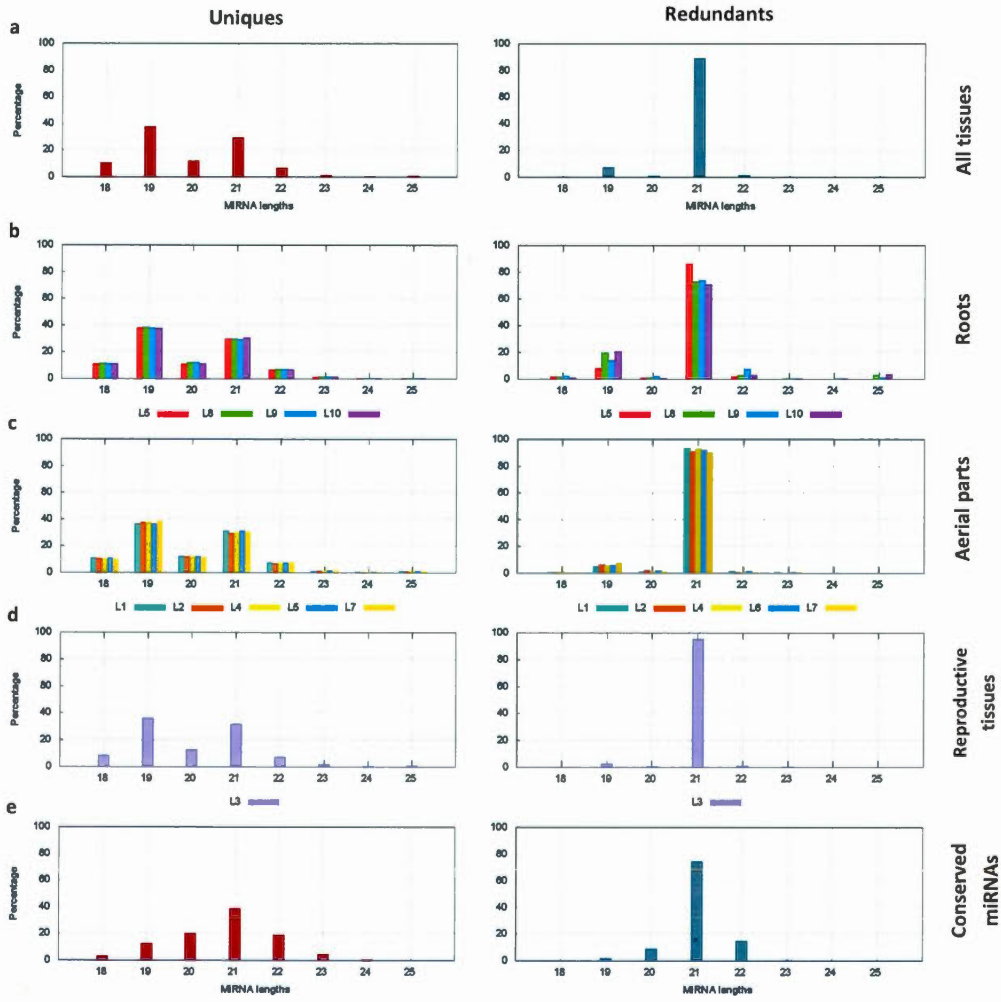


Figure B.7: miRNA length distribution in different tissues of hexaploid wheat. Percentages of unique (different sequences) or total redundant sequences were plotted according to the length of miRNAs. New miRNAs identified from a) all tissues (aerial parts. roots and reproductive tissues) from the ten libraries; b) roots (L5, L8, L9 and L10); c) aerial parts (L1, L2, L4, L6, L7); d) reproductive tissues (L3) independently of stress and genotypes; e) conserved miRNAs in the ten sequenced libraries identified using homology searches against miRBase. For detailed information about the libraries and conditions see Method B.1.1 and Table B.3.

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADN acide désoxyribonucléique. 7, 9, 10

ANN artificial neural networks. 1

ARN acide ribonucléique. 7, 8, 10

ARNi ARN interférent. 8

ARNm ARN messager. 7, 8, 119

ARNr ARN ribosomique. 7

ARNt ARN de transfert. 7, 8

EST expressed sequence tag. 76, 84, 116

HMM hidden Markov model. 2

miARN microARN. 8, 39, 42, 116–119

NGS Next-Generation Sequencing. 97, 99

RFLP restriction fragment length polymorphism. 10, 115

SVM support vector machines. 2

VHB virus de l'hépatite B. 115

VIH virus de l'immunodéficience humaine. 115

VPH virus du papillome humain. 115

GLOSSAIRE

apoptose Processus de la mort programmée d'une cellule. 8

endonucléase enzyme qui coupe au milieu d'un acide nucléique en fragments. 9

EST séquence partielle d'un ADN complémentaire, utilisée pour l'identification des gènes. 76

génomique étude du matériel génétique d'un individu ou d'une espèce codé dans son ADN ou ARN. 2

métagénomique étude du matériel génétique (ADN ou ARN), de plusieurs espèces, récupéré à partir d'échantillons d'environnements complexes (intestin, sols, océans, etc.). 2

polymère grande molécule constituée d'unités fondamentales appelées monomères reliées par des liaisons covalentes. 7

PubMed Base de données et moteur de recherche des méta-données bibliographiques de la littérature relative à la biologie et à la médecine. 38

transcriptome Ensemble des ARNs issus de la transcription du génome. 2

RÉFÉRENCES

Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V. et Vance, V. (2005). Computational prediction of miRNAs in arabidopsis thaliana. *Genome Research*, *5*.

Adams, J. et Rothman, E. (1982). Estimation of phylogenetic relationships from DNA restriction patterns and selection of endonuclease cleavage sites. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(11), 3560–3564.

Agharbaoui, Z., Leclercq, M., Remita, M. A., Badawi, M. A., Lord, E., Houde, M., Danyluk, J., Diallo, A. B. et Sarhan, F. (2015). An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress. *BMC Genomics*, *16*(1), 339.

Aha, D. W., Kibler, D. et Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66.

Alcantara, L. C. J., Cassol, S., Libin, P., Deforche, K., Pybus, O. G., Van Ranst, M., Galvao-Castro, B., Vandamme, A.-M. et de Oliveira, T. (2009). A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Research*, *37*(Web Server issue), W634–W642.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. et Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410.

Altschul, S. F., Madden, T. L., Schäffer, a. a., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M. *et al.* (2003). A uniform system for microRNA annotation. *RNA*, *9*(3), 277–279.

Badawi, M., Danyluk, J., Boucho, B., Houde, M. et Sarhan, F. (2007). The CBF gene family in hexaploid wheat and its relationship to the phylogenetic complexity of cereal CBFs. *Mol Genet Genomics*, *277*.

Badawi, M., Reddy, Y. V., Agharbaoui, Z., Tominaga, Y., Danyluk, J. et Sarhan, F. (2008). Structure and functional analysis of wheat ICE inducer of CBF expression genes. *Plant Cell Physiol*, *49*.

Bajla, I., Holländer, I., Fluch, S., Burg, K. et Kollár, M. (2005). An alternative method for electrophoretic gel image analysis in the GelMaster software. *Computer methods and programs in biomedicine*, *77*(3), 209–31.

Bao, Y., Chetvernin, V. et Tatusova, T. (2014). Improvements to pairwise sequence comparison (PASC) : a genome-based web tool for virus classification. *Archives of Virology*, *159*(12), 3293–3304.

Baumgartner, C., Bohm, C., Baumgartner, D., Marini, G., Weinberger, K., Olgemoller, B., Liebl, B. et Roscher, A. A. (2004). Supervised machine learning techniques for the classification of metabolic disorders in newborns. *Bioinformatics*, *20*(17), 2985–2996.

Ben-Bassat, M. (1982). 35 Use of distance measures, information measures and error bounds in feature evaluation. *Handbook of Statistics*, *2*, 773–791.

Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J Roy Stat Soc*, *85*.

Bernard, H.-U., Burk, R. D., Chen, Z., van Doorslaer, K., zur Hausen, H. et de Villiers, E. M. (2010). Classification of papillomaviruses (PVs) based on 189 pv types and proposal of taxonomic amendments. *Virology*, *401*(1), 70–79.

Bernard, H.-U., Chan, S.-Y., Manos, M. M., Ong, C.-K., Villa, L. L., Delius, H., Peyton, C. L., Bauer, H. M. et Wheeler, C. M. (1994). Identification and assessment of known and novel human papillomaviruses by polymerase chain reaction amplification, restriction fragment length polymorphisms, nucleotide sequence, and phylogenetic algorithms. *Journal of Infectious Diseases*, *170*(5), 1077–1085.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C. et Apweiler, R. (2009). QuickGO : a web-based tool for gene ontology searching. *Bioinformatics*, *25*(22), 3045–3046.

Blagus, R. et Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC bioinformatics*, *11*, 523.

Bonham-Carter, O., Steele, J. et Bastola, D. (2013). Alignment-free genetic sequence comparisons : a review of recent approaches by word analysis. *Briefings in bioinformatics*, *15*(6), 890–905.

Bonnet, E., He, Y., Billiau, K. et Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microrna targets, including target mimics. *Bioinformatics*, *26*(12), 1566–1568.

Boser, B. E., Guyon, I. M. et Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Dans *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152., San Mateo, CA. ACM Press.

Brady, A. et Salzberg, S. L. (2009). Phymm and PhymmBL : metagenomic phylogenetic classification with interpolated markov models. *Nature Methods*, *6*(9), 673–676.

Breakfield, N. W., Corcoran, D. L., Petricka, J. J., Shen, J., Sae-Seaw, J. et Rubio-Somoza, I. (2012). High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in arabidopsis. *Genome Research*, *22*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. et Stone, C. (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Breitbart, M. et Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, *13*(6), 278–284.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L., D'Amore, R. et Allen, A. M. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, *491*.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase ii promoter elements derived from 502 unrelated promoter sequences. *Journal of molecular biology*, *212*(4), 563–578.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. et Madden, T. L. (2009). BLAST+ : architecture and applications. *BMC bioinformatics*, *10*(1), 421.

Carollo, V., Matthews, D. E., Lazo, G. R., Blake, T. K., Hummel, D. D., Lui, N., Hane, D. L. et Anderson, O. D. (2005). GrainGenes 2.0. an improved resource for the small-grains community. *Plant physiology*, *139*(2), 643–651.

Carrington, J. C. et Ambros, V. (2003). Role of microRNAs in plant and animal development. *Science*, *301*.

Chang, C.-C. et Lin, C.-J. (2011). LIBSVM : A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, *2*(3), 27 :1–27 :27.

Chang, H.-W., Cheng, Y.-H., Chuang, L.-Y. et Yang, C.-H. (2010). SNP-RFLPing 2 : an updated and integrated PCR-RFLP tool for SNP genotyping. *BMC bioinformatics*, *11*, 173.

Chen, H.-M., Chen, L.-T., Patel, K., Li, Y.-H., Baulcombe, D. C. et Wu, S.-H. (2010). 22-nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proceedings of the National Academy of Sciences*, *107*(34), 15269–15274.

Childs, K. L., Hamilton, J. P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P. D., Town, C. D., Buell, C. R. et Chan, A. P. (2007). The TIGR plant transcript assemblies database. *Nucleic Acids Research*, *35*(suppl 1), D846–D851.

Colaiacovo, M., Lamontanara, A., Bernardo, L., Alberici, R., Crosatti, C. et Giusti, L. (2012). On the complexity of miRNA-mediated regulation in plants : novel insights into the genomic organization of plant miRNAs. *Biol Direct*, *7*.

Coordinators, N. R. (2016). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *44*(Database issue), D7–D19.

Cornuéjols, A. et Miclet, L. (2010). *Apprentissage artificiel : concepts et algorithmes* (2 éd.). Editions Eyrolles.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cover, T. et Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Cover, T. M. et Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

Cuperus, J. T., Fahlgren, N. et Carrington, J. C. (2011). Evolution and functional diversification of MIRNA genes. *The Plant Cell*, *23*(2), 431–442.

Daigle, B., Makarenkov, V. et Diallo, A. B. (2015). Effect of hundreds sequenced genomes on the classification of human papillomaviruses. In *Data Science, Learning by Latent Structures, and Knowledge Discovery* 309–318. Berlin, Heidelberg : Springer Berlin Heidelberg.

Darty, K., Denise, A. et Ponty, Y. (2009). VARNA : Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, *25*(15), 1974.

de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E. J., Wensing, A. M. J., van de Vijver, D. A., Boucher, C. A., Camacho, R. et Vandamme, A. M. (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, *21*(19), 3797–3800.

Degroeve, S., De Baets, B., Van de Peer, Y. et Rouzé, P. (2002). Feature subset selection for splice site prediction. *Bioinformatics*, *18*(suppl 2), S75–S83.

Dehury, B., Panda, D., Sahu, J., Sahu, M., Sarma, K., Barooah, M., Sen, P. et Modi, M. K. (2013). In silico identification and characterization of conserved miRNAs and their target genes in sweet potato (ipomoea batatas L.) expressed sequence tags (ESTs). *Plant signaling & behavior*, *8*(12).

Deng, P., Nie, X., Wang, L., Cui, L., Liu, P. et Tong, W. (2014). Computational identification and comparative analysis of miRNAs in wheat group 7 chromosomes. *Plant Mol Biol Rep*, *32*.

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K. et Nattkemper, T. W. (2009). TACOA : taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics*, *10*, 56.

Dinger, M. E., Pang, K. C., Mercer, T. R. et Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA : challenges and ambiguities. *PLoS Comput Biol*, *4*(11), e1000176.

Dryanova, A., Zakharov, A. et Gulick, P. J. (2008). Data mining for miRNAs and their targets in the Triticeae. *Genome*, *51*(6), 433–443.

Duda, R. O., Hart, P. E. et Stork, D. G. (2001). *Pattern classification* (2 éd.). John Wiley & Sons.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461.

Ester, M., Kriegel, H.-P., Sander, J. et Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Dans *Kdd*, volume 96, 226–231.

Fahlgren, N., Jogdeo, S., Kasschau, K. D., Sullivan, C. M., Chapman, E. J. et Laubinger, S. (2010). Microrna gene evolution in Arabidopsis lyrata and arabidopsis thaliana. *Plant Cell*, *22*.

Felsenstein, J. (1992). Phylogenies from restriction sites : A maximum-likelihood approach. *Evolution*, *46*(1), 159–173.

Freund, Y. et Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S. et Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, *26*(4), 407–415.

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W. et Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microrna genes in seven animal clades. *Nucleic Acids Research*, *40*(1), 37–52.

Gao, F., Robertson, D. L., Carruthers, C. D., Morrison, S. G., Jian, B., Chen, Y., Barré-Sinoussi, F., Girard, M., Srinivasan, A., Abimiku, A., Shaw, G., Sharp, P. et Hahn, B. (1998). A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *Journal of Virology*, *72*(7), 5680–5698.

Gentile, A., Ferreira, T. H., Mattos, R. S., Dias, L. I., Hoshino, A. A., Carneiro, M. S., Souza, G. M., Calsa Jr, T., Nogueira, R. M., Endres, L. *et al.* (2013). Effects of drought on the microtranscriptome of field-grown sugarcane plants. *Planta*, *237*(3), 783–798.

Gish, W., States, D. J. *et al.* (1993). Identification of protein coding regions by database similarity search. *Nature genetics*, *3*(3), 266–272.

Gkirtzou, K., Tsamardinos, I., Tsakalides, P. et Poirazi, P. (2010). MatureBayes : a probabilistic algorithm for identifying the mature mirna within novel precursors. *PLoS One*, *5*(8), e11843.

Guan, D.-G., Liao, J.-Y., Qu, Z.-h., Zhang, Y. et Qu, L.-h. (2011). mirExplorer : detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features. *RNA biology*, *8*(5), 922–934.

Gudyś, A., Szcześniak, M. W., Sikora, M. et Makałowska, I. (2013). HuntMi : an efficient and taxon-specific approach in pre-miRNA identification. *BMC bioinformatics*, *14*, 83.

Guo, S., Xu, Y., Liu, H., Mao, Z., Zhang, C. et Ma, Y. (2013). The interaction between OsMADS57 and OsTB1 modulates rice tillering via DWARF14. *Nat Commun*, *4*.

Gutierrez, L., Mongelard, G., Floková, K., Pacurar, D. I., Novák, O. et Staswick, P. (2012). Auxin controls Arabidopsis adventitious root initiation by regulating jasmonic acid homeostasis. *Plant Cell*, *6*.

Hackenberg, M., Rodríguez-Ezpeleta, N. et Aransay, A. M. (2011). miRanalyzer : an update on the detection and analysis of micrornas in high-throughput sequencing experiments. *Nucleic Acids Research*, *39*(suppl 2), W132–W138.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J. M. et Aransay, A. M. (2009). miRanalyzer : a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research*, *37*(suppl 2), W68–W76.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. et Witten, I. H. (2009). The WEKA data mining software : An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hamel, F., Breton, C. et Houde, M. (1998). Isolation and characterization of wheat aluminum-regulated genes : possible involvement of aluminum as a pathogenesis response elicitor. *Planta*, *205*(4), 531–538.

Han, J., Kamber, M. et Pei, J. (2011). *Data mining : concepts and techniques* (3 éd.). Elsevier.

Han, J., Kong, M., Xie, H., Sun, Q., Nan, Z., Zhang, Q. et Pan, J. (2013). Identification of miRNAs and their targets in wheat (Triticum aestivum L.) by EST analysis. *Genetics and Molecular Research*, *12*(3), 3793.

Han, R., Jian, C., Lv, J., Yan, Y., Chi, Q., Li, Z., Wang, Q., Zhang, J., Liu, X. et Zhao, H. (2014). Identification and characterization of microRNAs in the flag leaf and developing seed of wheat (Triticum aestivum L.). *BMC Genomics*, *15*(1), 289.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A. *et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318.

Helvik, S. A., Snøve, O. et Sætrom, P. (2007). Reliable prediction of drosha processing sites improves microRNA gene prediction. *Bioinformatics*, *23*(2), 142–149.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, *31*.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* Cambridge, MA, USA : MIT Press.

Hossain Khan, M. S., Tawaraya, K., Sekimoto, H., Koyama, H., Kobayashi, Y., Murayama, T., Chuba, M., Kambayashi, M., Shiono, Y., Uemura, M. *et al.* (2009). Relative abundance of δ5-sterols in plasma membrane lipids of root-tip cells correlates with aluminum tolerance of rice. *Physiologia plantarum*, *135*(1), 73–83.

Houde, M., Belcaid, M., Ouellet, F., Danyluk, J., Monroy, A. F., Dryanova, A., Gulick, P., Bergeron, A., Laroche, A., Links, M. G. *et al.* (2006). Wheat EST resources for functional genomics of abiotic stress. *BMC Genomics*, *7*(1), 149.

Houde, M. et Diallo, A. O. (2008). Identification of genes and pathways associated with aluminum stress and tolerance using transcriptome profiling of wheat near-isogenic lines. *BMC Genomics*, *9*(1), 400.

Hsieh, L. C., Lin, S. I., Shih, A. C., Chen, J. W., Lin, W. Y. et Tseng, C. Y. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in arabidopsis by deep sequencing. *Plant Physiol*, *151*.

Huang, T.-H., Fan, B., Rothschild, M. F., Hu, Z.-L., Li, K. et Zhao, S.-H. (2007). MiRFinder : an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC bioinformatics*, *8*, 341.

Hunt, E. B., Marin, J. et Stone, P. J. (1966). *Experiments in induction.* New York : Academic Press.

Huson, D. H., Auch, A. F., Qi, J. et Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386.

Jain, A. K. (2010). Data clustering : 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651–666.

Jain, M., Ghanashyam, C. et Bhattacharjee, A. (2010). Comprehensive expression analysis suggests overlapping and specific roles of rice glutathione S-transferase genes during development and stress responses. *BMC Genomics*, *11*(1), 73.

Janini, L. M., Pieniazek, D., Peralta, J. M., Schechter, M., Tanuri, A., Vicente, A. C. P., dela Torre, N., Pieniazek, N. J., Luo, C.-C., Kalish, M. L., Schochetman, G. et Rayfield, M. a. (1996). Identification of single and dual infections with distinct subtypes of human immunodeficiency virus type 1 by using restriction fragment length polymorphism analysis. *Virus Genes*, *13*(1), 69–81.

Janska, A., Marsik, P., Zelenkova, S. et Ovesna, J. (2010). Cold stress and acclimation : what is important for metabolic adjustment? *Plant Biol (Stuttg)*, *12*.

Jensen, L. J. et Bateman, A. (2011). The rise and fall of supervised machine learning techniques. *Bioinformatics*, *27*(24), 3331–3332.

Jeong, D.-H., Park, S., Zhai, J., Gurazada, S. G. R., De Paoli, E., Meyers, B. C. et Green, P. J. (2011). Massive analysis of rice small RNAs : mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. *The Plant Cell*, *23*(12), 4185–4207.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. et Lu, Z. (2007). MiPred : classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, *35*(suppl 2), W339–W344.

Jin, W., Li, N., Zhang, B., Wu, F., Li, W., Guo, A. et Deng, Z. (2008). Identification and verification of microRNA in wheat (Triticum aestivum). *J Plant Res*, *121*(3), 351–355.

John, G. H. et Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. Dans *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, 338–345., San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jones-Rhoades, M. W. et Bartel, D. P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Molecular cell*, *14*(6), 787–799.

Jones-Rhoades, M. W., Bartel, D. P. et Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol*, *57*, 19–53.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. et Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, *110*.

Kadri, S., Hinman, V. et Benos, P. V. (2009). HHMMiR : efficient de novo prediction of microRNAs using hierarchical hidden markov models. *BMC bioinformatics*, *10*(Suppl 1), S35.

Kal, A. J., Zonneveld, A. J., Benes, V., Vandenberg, M., Koerkamp, M. G. et Albermann, K. (1999). Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol Biol Cell*, *10*.

Kantar, M., Akpinar, B. A., Valárik, M., Lucas, S. J., Doležel, J. et Hernández, P. (2012). Subgenomic analysis of microRNAs in polyploid wheat. *Funct Integr Genomics*, *12*.

Kelley, D. R. et Salzberg, S. L. (2010). Clustering metagenomic sequences with interpolated markov models. *BMC Bioinformatics*, *11*(1), 544.

Kim, J., Ahn, Y., Lee, K., Park, S. H. et Kim, S. (2010). A classification approach for genotyping viral sequences based on multidimensional scaling and linear discriminant analysis. *BMC bioinformatics*, *11*, 434.

Kim, J. S., Kim, K. A., Oh, T. R., Park, C. M. et Kang, H. (2008). Functional characterization of DEAD-box rna helicases in arabidopsis thaliana under abiotic stress conditions. *Plant Cell Physiol*, *49*.

Kozomara, A. et Griffiths-Jones, S. (2011). miRBase : integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, *39*(suppl 1), D152–D157.

Kozomara, A. et Griffiths-Jones, S. (2014). miRBase : annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, *42*.

Kurepin, L. V., Dahal, K. P., Savitch, L. V., Singh, J., Bode, R. et Ivanov, A. G. (2013). Role of CBFs as integrators of chloroplast redox, phytochrome and plant hormone signaling during cold acclimation. *Int J Mol Sci*, *14*.

Kurtoglu, K. Y., Kantar, M., Lucas, S. J. et Budak, H. (2013). Unique and conserved microRNAs in wheat chromosome 5D revealed by next-generation sequencing. *PLoS One*, *8*.

Langley, P., Iba, and, W. et Thompson, K. (1992). An analysis of bayesian classifiers. Dans *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, 223–228., Menlo Park, CA. AAAI Press.

Lauber, C. et Gorbalenya, A. E. (2012). Partitioning the genetic diversity of a virus family : Approach and evaluation through a case study of picornaviruses. *Journal of Virology*, *86*(7), 3890–3904.

Leclercq, M., Diallo, A. B. et Blanchette, M. (2013). Computational prediction of the localization of microRNAs within their pre-miRNA. *Nucleic Acids Research*, *41*(15), 7200–7211.

Li, A., Liu, D., Wu, J., Zhao, X., Hao, M. et Geng, S. (2014). mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. *Plant Cell*, *26*.

Li, X., Hongwu, B., Dafeng, S., Shengyun, M., Ning, H. et Junhui, W. (2013a). Flowering time control in ornamental gloxinia (Sinningia speciosa). *Ann Bot*, *111*.

Li, Y.-F., Zheng, Y., Addo-Quaye, C., Zhang, L., Saini, A., Jagadeeswaran, G., Axtell, M. J., Zhang, W. et Sunkar, R. (2010). Transcriptome-wide identification of microRNA targets in rice. *The Plant Journal*, *62*(5), 742–759.

Li, Y.-F., Zheng, Y., Jagadeeswaran, G. et Sunkar, R. (2013b). Characterization of small RNAs and their target genes in wheat seedlings using sequencing-based approaches. *Plant Science*, *203*, 17–24.

Liang, G. et Yu, D. (2010). Reciprocal regulation among miR395, APS and SULTR2,1 in Arabidopsis thaliana. *Plant Signal Behav*, *10*.

Libbrecht, M. W. et Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332.

Lin, W.-J. et Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, *14*(1), 13–26.

Ling, H. Q., Zhao, S., Liu, D., Wang, J., Sun, H. et Zhang, C. (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, *496*.

Ling, K.-H., Brautigan, P. J., Hahn, C. N., Daish, T., Rayner, J. R., Cheah, P.-S., Raison, J. M., Piltz, S., Mann, J. R., Mattiske, D. M. *et al.* (2011). Deep sequencing analysis of the developing mouse brain reveals a novel microRNA. *BMC Genomics*, *12*(1), 1.

Liu, Z., Meng, J. et Sun, X. (2008). A novel feature-based method for whole genome phylogenetic analysis without alignment : Application to HEV genotyping and subtyping. *Biochemical and Biophysical Research Communications*, *368*(2), 223–230.

Lord, E., Leclercq, M., Boc, A., Diallo, A. B. et Makarenkov, V. (2012). Armadillo 1.1 : an original workflow platform for designing and conducting phylogenetic analysis and simulations. *PLoS One*, *7*(1), e29903.

Lucas, S. J. et Budak, H. (2012). Sorting the wheat from the chaff : identifying miRNAs in genomic survey sequences of Triticum aestivum chromosome 1AL. *PLoS One*, *7*(7), e40859.

Mantaci, S., Restivo, A. et Sciortino, M. (2008). Distance measures for biological sequences : Some recent approaches. *International Journal of Approximate Reasoning*, *47*(1), 109–124.

Maramis, C. F., Delopoulos, A. N. et Lambropoulos, A. F. (2011). A computerized methodology for improved virus typing by PCR-RFLP gel electrophoresis. *IEEE Transactions on Biomedical Engineering*, *58*(8), 2339–2351.

Marrs, K. A. (1996). The functions and regulation of glutathione S-transferases in plants. *Annual Review of Plant Biology*, *47*(1), 127–158.

Matsen, F. A., Kodner, R. B. et Armbrust, E. V. (2010). pplacer : linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538.

Meng, F., Liu, H., Wang, K., Liu, L., Wang, S., Zhao, Y., Yin, J. et Li, Y. (2013). Development-associated microRNAs in grains of wheat (Triticum aestivum l.). *BMC Plant Biology*, *13*(1), 140.

Menossi, M. M., Gentile, A., Ferreira, T. H., Mattos, R. et Dias, L. I. (2015). MicroRNAs and drought responses in sugarcane. *Frontiers in Plant Science*, *6*, 58.

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., Cao, X., Carrington, J. C., Chen, X., Green, P. J. *et al.* (2008). Criteria for annotation of plant MicroRNAs. *The Plant Cell*, *20*(12), 3186–3190.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mizokami, M., Nakano, T., Orito, E., Tanaka, Y., Sakugawa, H., Mukaide, M. et Robertson, B. H. (1999). Hepatitis B virus genotype assignment using restriction fragment length polymorphism patterns. *FEBS Letters*, *450*(1-2), 66–71.

Muñoz, N., Bosch, F. X., de Sanjosé, S., Herrero, R., Castellsagué, X., Shah, K. V., Snijders, P. J. et Meijer, C. J. (2003). Epidemiologic classification of human papillomavirus types associated with cervical cancer. *New England Journal of Medicine*, *348*(6), 518–527.

Murtas, G., Reeves, P. H., Fu, Y. . F., Bancroft, I., Dean, C. et Coupland, G. (2003). A nuclear protease required for flowering time regulation in Arabidopsis reduces the abundance of small ubiquitin-related modifier conjugates. *Plant Cell*, *15*.

Nakao, T., Enomoto, N., Takada, N., Takada, A. et Date, T. (1991). Typing of hepatitis C virus genomes by restriction fragment length polymorphism. *Journal of General Virology*, *72*(9), 2105–2112.

Nam, J.-W., Shin, K.-R., Han, J., Lee, Y., Kim, V. N. et Zhang, B.-T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research*, *33*(11), 3570–81.

Nobre, R. J., de Almeida, L. P. et Martins, T. C. (2008). Complete genotyping of mucosal human papillomavirus using a restriction fragment length polymorphism analysis and an original typing algorithm. *Journal of Clinical Virology*, *42*(1), 13–21.

Ohler, U., Liao, G.-c., Niemann, H. et Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. *Genome biology*, *3*(12), 1.

Ondov, B. D., Varadarajan, A., Passalacqua, K. D. et Bergman, N. H. (2008). Efficient mapping of applied biosystems SOLiD sequence data to a reference genome for functional genomics applications. *Bioinformatics*, *24*.

Osman, I. H. et Kelly, J. P. (dir.) (1996). *Meta-heuristics : theory and applications* (1 éd.). Springer US.

Oulas, A., Boutla, A., Gkirtzou, K., Reczko, M., Kalantidis, K. et Poirazi, P. (2009). Prediction of novel microRNA genes in cancer-associated genomic regions–a combined computational and experimental approach. *Nucleic Acids Research*, *37*(10), 3276–87.

Pandey, B., Gupta, O. P., Pandey, D. M., Sharma, I. et Sharma, P. (2013). Identification of new stress-induced microRNA and their targets in wheat using computational approach. *Plant Signal Behav*, *8*.

Pandey, R., Joshi, G., Bhardwaj, A. R., Agarwal, M. et Katiyar-Agarwal, S. (2014). A comprehensive genome-wide study on tissue-specific and abiotic stress-specific miRNAs in Triticum aestivum. *PLoS One*, *9*(4), e95800.

Paolacci, A. R., Tanzarella, O. A., Porceddu, E., Varotto, S. et Ciaffi, M. (2007). Molecular and phylogenetic analysis of MADS-box genes of MIKC type and chromosome location of SEP-like genes in wheat (Triticum aestivum L.). *Mol Genet Genomics*, *278*.

Patil, K. R., Haider, P., Pope, P. B., Turnbaugh, P. J., Morrison, M., Scheffer, T. et McHardy, A. C. (2011). Taxonomic metagenome sequence assignment with structured output models. *Nature Methods*, *8*(3), 191–192.

Perz, J. F., Armstrong, G. L., Farrington, L. A., Hutin, Y. J. et Bell, B. P. (2006). The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *Journal of Hepatology*, *45*(4), 529–538.

Pevzner, P. (2000). *Computational molecular biology : an algorithmic approach.* Cambridge, MA : MIT press.

Picardi, E. et Pesole, G. (2010). Computational methods for ab initio and comparative gene finding. *Data Mining Techniques for the Life Sciences*, 269–284.

Piriyapongsa, J. et Jordan, I. K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA*, *14*.

Pirooznia, M., Yang, J. Y., Yang, M. Q. et Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, *9*(Suppl 1), S13.

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D. L. et Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine*, *15*(8), 871–872.

Pond, S. L. K., Posada, D., Stawiski, E., Chappey, C., Poon, A. F. Y., Hughes, G., Fearnhill, E., Gravenor, M. B., Brown, A. J. L. et Frost, S. D. W. (2009). An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput Biol*, *5*(11).

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (dir.), *Expert Systems in the Microelectronic Age*. Edinburgh, UK : Edinburgh University Press.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Rajagopalan, R., Vaucheret, H., Trejo, J. et Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microrna in Arabidopsis thaliana. *Genes Dev*, *20*.

Reeves, P. H., Murtas, G., Dash, S. et Coupland, G. (2002). Early in short days 4, a mutation in Arabidopsis that causes early flowering and reduces the mRNA abundance of the floral repressor FLC. *Development*, *129*.

Ribeiro-dos Santos, Â., Khayat, A., Silva, A., Alencar, D., Lobato, J. et Luz, L. (2010). Ultra-deep sequencing reveals the microRNAs expression pattern of the human stomach. *PLoS One*, *5*.

Roberts, R. J., Vincze, T., Posfai, J. et Macelis, D. (2015). REBASE–a database for DNA restriction and modification : enzymes, genes and genomes. *Nucleic Acids Research*, *43*(Database issue), D298–D299.

Robertson, D., Anderson, J., Bradac, J., Carr, J., Foley, B., Funkhouser, R., Gao, F., Hahn, B., Kalish, M., Kuiken, C., Learn, G., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P., Wolinsky, S. et B., K. (2000). HIV-1 nomenclature proposal. *Science*, *288*(5463), 55–56.

Rogers, K. et Chen, X. (2013). Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell*, *25*.

Rosen, G. L., Reichenberger, E. R. et Rosenfeld, A. M. (2011). NBC : the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, *27*(1), 127–129.

Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rousseeuw, P. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing : Explorations in the Microstructure of Cognition, Vol. 1* 318–362. Cambridge, MA, USA : MIT Press.

Salzberg, S. L., Delcher, A. L., Kasif, S. et White, O. (1998). Microbial gene identification using interpolated markov models. *Nucleic Acids Research*, *26*(2), 544–548.

Schaefer, S. (2007). Hepatitis B virus taxonomy and hepatitis B virus genotypes. *World Journal of Gastroenterology*, *13*(1), 14–21.

Schulte, J., Marschall, T., Martin, M., Rosenstie, P., Mestdagh, P. et Schlierf, S. (2010). Deep sequencing reveals differential expression of micrornas in favorable versus unfavorable neuroblastoma. *Nucleic Acids Research*, *38*.

Shaik, R. et Ramakrishna, W. (2012). Bioinformatic analysis of epigenetic and microRNA mediated regulation of drought responsive genes in rice. *PLoS One*, *7*(11), e49331.

Smit, A. F. A., Hubley, R. et Green, P. (2010). *RepeatMasker*.

Struck, D., Lawyer, G., Ternes, A.-M., Schmit, J.-C. et Bercoff, D. P. (2014). COMET : adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Research*, *42*(18), e144.

Sun, F., Guo, G., Du, J., Guo, W., Peng, H., Ni, Z., Sun, Q. et Yao, Y. (2014). Whole-genome discovery of miRNAs and their targets in wheat (Triticum aestivum L.). *BMC Plant Biology*, *14*(1), 142.

Sun, G., Stewart, C. N., Xiao, P. et Zhang, B. (2012). MicroRNA expression analysis in the cellulosic biofuel crop Switchgrass panicum virgatum under abiotic stress. *PLoS One*, *7*.

Sunkar, R., Girke, T., Jain, P. K. et Zhu, J. K. (2005). Cloning and characterization of microRNAs from rice. *Plant Cell*, *17*.

Sunkar, R., Li, Y. F. et Jagadeeswaran, G. (2012). Functions of microRNAs in plant stress responses. *Trends Plant Sci*, *4*.

Szarzynska, B., Sobkowiak, L., Pant, B. D., Balazadeh, S., Scheible, W. R. et Mueller-Roeber, B. (2009). Gene structures and processing of Arabidopsis thaliana HYL1-dependent pri-miRNAs. *Nucleic Acids Research*, *37*.

Tan, P.-N., Steinbach, M. et Kumar, V. (2006). *Introduction to data mining*. Pearson Addison Wesley Boston.

Tang, Z., Zhang, L., Xu, C., Yuan, S., Zhang, F., Zheng, Y. et Zhao, C. (2012). Uncovering small RNA-mediated responses to cold stress in a wheat thermosensitive genic male-sterile line by deep sequencing. *Plant Physiology*, *159*(2), 721–738.

Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R. et Draghici, S. (2007). Machine learning and its applications to biology. *PLoS Comput Biol*, *3*(6), e116.

Templeton, A. R. (1983). Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of human and the apes. *Evolution*, *37*(2), 221–244.

Tzahor, S., Man-Aharonovich, D., Kirkup, B. C., Yogev, T., Berman-Frank, I., Polz, M. F., Béjà, O. et Mandel-Gutfreund, Y. (2009). A supervised learning approach for taxonomic classification of core-photosystem-ii genes and transcripts in the marine environment. *BMC Genomics*, *10*, 229.

Van Belkum, A., Struelens, M., de Visser, A., Verbrugh, H. et Tibayrenc, M. (2001). Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. *Clinical microbiology reviews*, *14*(3), 547–560.

Vapnik, V. N. et Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, *16*(2), 264–280.

Várallyay, E., Burgyán, J. et Havelda, Z. (2008). MicroRNA detection by northern blotting using locked nucleic acid probes. *Nat Protoc*, *3*.

Vinga, S. et Almeida, J. (2003). Alignment-free sequence comparison–a review. *Bioinformatics*, *19*(4), 513–523.

Wagatsuma, T., Ishikawa, S., Uemura, M., Mitsuhashi, W., Kawamura, T., Khan, S. H. et Tawaraya, K. (2005). Plasma membrane lipids are the powerful components for early stage aluminum tolerance in triticale. *Soil Science & Plant Nutrition*, *51*(5), 701–704.

Wang, B., Sun, Y., Song, N., Wang, X., Feng, H., Huang, L. et Kang, Z. (2013). Identification of UV-B-induced microRNAs in wheat. *Genetics and Molecular Research*, *12*(4), 4213.

Wang, J. W., Wang, L. J., Mao, Y. B., Cai, W. J., Xue, H. W. et Chen, X. Y. (2005). Control of root cap formation by microRNA-targeted auxin response factors in arabidopsis. *Plant Cell*, *17*.

Wang, L., Huang, H., Fan, Y., Kong, B., Hu, H. et Hu, K. (2014). Effects of downregulation of microRNA-181a on H2O2-induced H9c2 cell apoptosis via the mitochondrial apoptotic pathway. *Oxid Med Cell Longev*, *2014*.

Wang, Q., Garrity, G. M., Tiedje, J. M. et Cole, J. R. (2007). Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267.

Wei, B., Cai, T., Zhang, R., Li, A., Huo, N., Li, S., Gu, Y. Q., Vogel, J., Jia, J., Qi, Y. *et al.* (2009). Novel microRNAs uncovered by deep sequencing of small RNA transcriptomes in bread wheat (Triticum aestivum L.) and Brachypodium distachyon (L.) Beauv. *Functional & integrative genomics*, *9*(4), 499–511.

Williams, R. C. (1989). Restriction fragment length polymorphism (RFLP). *American Journal of Physical Anthropology*, *32*(S10), 159–184.

Witten, I., Frank, E. et Hall, M. (2011). *Data Mining : Practical Machine Learning Tools and Techniques* (3 éd.). The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Xin, M., Wang, Y., Yao, Y., Song, N., Hu, Z. et Qin, D. (2011). Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and sbs sequencing. *BMC Plant Biology*, *11*.

Xing, S., Salinas, M., Höhmann, S., Berndtgen, R. et Huijser, P. (2011). miR156-targeted and nontargeted SBP-box transcription factors act in concert to secure male fertility in Arabidopsis. *Plant Cell*, *22*.

Xing, Z., Pei, J. et Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, *12*(1), 40–48.

Xu, L., Wang, Y., Zhai, L., Xu, Y., Wang, L., Zhu, X., Gong, Y., Yu, R., Limera, C. et Liu, L. (2013). Genome-wide identification and characterization of cadmium-responsive microRNAs and their target genes in radish (Raphanus sativus L.) roots. *Journal of experimental botany*, p. ert240.

Xue, C., Li, F., He, T., Liu, G.-P., Li, Y. et Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics*, *6*, 310.

Yao, Y., Guo, G., Ni, Z., Sunkar, R., Du, J. et Zhu, J. K. (2007). Cloning and characterization of microRNAs from wheat Triticum aestivum L. *Genome Biology*, *8*.

Yin, Z. et Shen, F. (2010). Identification and characterization of conserved microRNAs and their target genes in wheat (Triticum aestivum). *Genet Mol Res*, *9*(2), 1186–1196.

Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L. C. et Showe, M. K. (2006). Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, *22*(11), 1325–34.

Yu, C., Hernandez, T., Zheng, H., Yau, S.-C., Huang, H.-H., He, R. L., Yang, J. et Yau, S. S.-T. (2013). Real time classification of viruses in 12 dimensions. *PLoS One*, *8*(5), e64328.

Zeng, Q.-Y., Yang, C.-Y., Ma, Q.-B., Li, X.-P., Dong, W.-W. et Nian, H. (2012). Identification of wild soybean miRNAs and their target genes responsive to aluminum stress. *BMC Plant Biology*, *12*(1), 182.

Zhang, B. H., Pan, X. P., Wang, Q. L., Cobb, G. P. et Anderson, T. A. (2005). Identification and characterization of new plant microrna using EST analysis. *Cell Res*, *15*.

Zhang, Z., Yu, J., Li, D., Liu, F. et Zhou, X. (2010). PMRD : plant microRNA database. *Nucleic Acids Research*, *38*.

Zhao, X., Liu, X., Guo, C., Gu, J. et Kai, X. (2013). Identification and characterization of microRNAs from wheat Triticum aestivum L. under phosphorus deprivation. J. *Plant Biochem Biotechnol*, *22*.

Zhou, J., Foster, D. P., Stine, R. a. et Ungar, L. H. (2006). Streamwise feature selection. *Journal of Machine Learning Research*, *7*, 1861—1885.

Zhou, X., Ruan, J., Wang, G. et Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. *PLoS computational biology*, *3*(3), e37.