Application de techniques de fouille de données en Bio-informatique

Gérard Ramstein

Ecole Centrale de Nantes Université de Nantes Ecole des Mines de Nantes

Ecole Doctorale STIM

Sciences et Technologies de l'Information et des Matériaux

Année 2012

N° E.D. : 503

HABILITATION A DIRIGER DES RECHERCHES
Spécialité : INFORMATIQUE

Présentée par

Gérard Ramstein

le 7 juin 2012
à l'Ecole polytechnique de l'université de Nantes

Application de techniques de fouille de données en Bio-informatique

Jury

Rapporteurs :	Anne-Claude Camproux, Professeur à l'université Paris Diderot
	Robin Gras, Professeur associé à l'université de Windsor, Canada
	Alain Guénoche, Directeur de recherche CNRS, IML, Marseille
Examinateurs :	Jin-Kao Hao, Professeur à l'université d'Angers
	Rémi Houlgatte, Directeur de Recherche, INSERM U954, Nancy
	Pascale Kuntz-Cosperec, Professeur à l'université de Nantes

Directrice de HdR : Pascale Kuntz-Cosperec
Laboratoire : LINA (Laboratoire d'Informatique de Nantes Atlantique)

Remerciements

Je tiens à adresser mes plus vifs remerciements aux membres du jury. Je remercie Anne-Claude Camproux, Robin Gras et Alain Guénoche pour avoir accepté d'être mes rapporteurs, ainsi que Jin-Kao Hao, Rémi Houlgatte et Pascale Kuntz-Cosperec pour l'intérêt porté à mes travaux.

Les travaux décrits dans ce mémoire ont été effectués au sein de l'équipe Connaissances et Décision du Laboratoire Informatique de Nantes-Atlantique. Je suis très reconnaissant envers son responsable Pascale Kuntz-Cosperec pour son soutien dans mes recherches. J'adresse à Pascale et ainsi qu'à tous les membres de cette chaleureuse équipe un grand merci pour les excellentes conditions de travail et l'ambiance stimulante qu'ils ont su développées. J'adresse tout particulièrement ces remerciements à Henri Briand, responsable de l'équipe lors de mon arrivée au LINA, pour m'avoir accordé sa confiance et pour m'avoir incité à ouvrir un champ de recherche émergent et épanouissant. J'exprime également toute ma gratitude envers Régis Gras, qui m'a permis d'aborder l'analyse bio-informatique sous l'angle de l'Analyse Statistique Implicative.

Mes travaux doivent énormément aux fructueuses collaborations entretenues avec les équipes INSERM nantaises. Je dois beaucoup à Yannick Jacques, responsable de l'équipe Cytokines et Récepteurs en Immuno-Cancérologie (unité Inserm U892, Centre de Recherche en Cancérologie Nantes-Angers). Yannick m'a permis de faire mes premiers pas en analyse de séquences biologiques, en m'apportant sa science sur de nombreux points qui me seraient sinon à jamais demeurés obscurs. De même, sans l'expérience de Jean Léger, alors responsable de l'équipe Génomique dans l'unité INSERM U533, je n'aurais pas pu appréhender aussi subtilement les arcanes des puces à ADN. Son successeur, Rémi Houlgatte m'a ensuite confié l'animation de la plate-forme de Bio-informatique de Nantes et je lui en suis infiniment reconnaissant. Grâce à lui, mes compétences en matière d'analyse du transcriptome se sont profondément enrichies. Je remercie son équipe et les membres de la plate-forme pour la qualité de leur accueil et les fructueuses discussions entretenues. Ils m'ont apporté une vision plus biologique des problématiques et ouvert de nouvelles perspectives de recherche. Ma gratitude va à Daniel Baron, post-doctorant de l'équipe pour ses conseils avisés, ainsi qu'aux ingénieurs de la plate-forme de bio-informatique : Audrey Bihouée, Raluca Teusan, Emeric Dubois et Edouard Hirchaud.

J'adresse mes remerciements aux doctorants que j'ai encadrés. Nicolas Beaume, Julien Lorec, Jérôme Mikolajczak et Hoai Tuong Nguyen ont bien voulu se confronter à mes propositions de recherche et me faire confiance dans la poursuite de leurs travaux.

Je suis enfin reconnaissant envers mes collègues enseignants de Polytech Nantes, avec lesquels je partage de nombreux moments conviviaux dans un environnement privilégié et stimulant.

Table des matières

Liste des figures

Liste des tables

Notations

Notations relatives aux séquences protéiques

$\mathcal{M}$	ensemble de motifs communs à une famille de protéines
X	espace d'entrée d'un classifieur SVM
f	fonction du classifieur SVM associant à une entrée une valeur réelle.
h	fonction d'étiquetage du SVM
Φ	fonction de projection de X vers l'espace des caractéristiques
$\mathcal{F}$	espace des caractéristiques du SVM
K	fonction noyau du SVM
Ω	alphabet des acides aminés
$\Gamma(\Omega)$	alphabet des classes d'acides aminés
$S(\Omega)$	ensemble des singletons obtenus à partir de Ω
$C(\Omega)$	ensemble des classes physico-chimiques de Taylor
m	un motif obtenu à partir de l'alphabet $\Gamma(\Omega)$
c(m)	fonction de coût d'un motif
sp(m)	spécificité d'un motif
Ψ	ensemble de séquences protéiques
pbest	meilleure position rencontrée pour une particule
gbest	meilleure position rencontrée pour un essaim de particules
$v_i(t)$	vecteur de vitesse d'une particule i à l'instant t
$x_i(t)$	position d'une particule i à l'instant t
ϕ	facteur d'ajustement d'un programme π
π	programme généré par un essaim grammatical
$AUC(\pi)$	aire sous la courbe ROC pour un programme π

Notations relatives aux données de puces à ADN

M	matrice de données d'expression
$pr(k)$	profil d'expression d'un gène k
$rang(pr(k))$	séquence des observations obtenue à partir de $pr(k)$
$r_k(i)$	intervalle de rang représentant une sous-séquence de $rang(pr(k))$
$\mathcal{A}$	jeu d'apprentissage
$\mathcal{G}$	ensemble de gènes
O	ensemble des observations
C	ensemble des classes d'observations
$\mathcal{L}$	fonction d'étiquetage qui associe à une observation une classe
$\mathcal{R}$	ensemble de règles
φ (A → B)	intensité d'implication d'une règle

Avant-propos

Ce mémoire est le fruit de mes recherches au sein de l'équipe Connaissances et Décision du LINA. Depuis mon intégration en 2000 dans cette équipe, mes recherches ont porté sur des applications de techniques de fouille de données en bio-informatique. Ce thème était alors émergent, tout particulièrement en France et même à l'étranger : la revue *BMC Bioinformatics* a sorti son premier volume la dernière année du deuxième millénaire. L'analyse de séquences occupait alors le devant de la scène, probablement parce que le traitement de chaînes de caractères apparaissait comme un domaine d'application immédiat pour un informaticien déconcerté par la complexité de la biologie moléculaire. L'analyse de données de puces à ADN était alors encore balbutiante : l'étude célèbre de Golub *et al.* portant sur deux types de leucémies venait d'être publiée [Golub 1999] et l'utilisation de la fouille de données encore peu considérée. Le terme *data mining*, accolé à celui de puces à ADN, est encore peu employé. Un article l'utilisera un peu plus tard pour décrire les défis que pose ce champ d'investigation sous le titre : *Microarray data mining: facing the challenges* [Piatetsky-Shapiro 2003].

Je me suis intéressé de manière conjointe à deux axes de recherche en bio-informatique : d'une part à la recherche d'homologues distants dans des familles de protéines, et d'autre part à l'analyse de données de puces à ADN. Ces deux thèmes sont liés à des collaborations avec des équipes INSERM nantaises, et ont été abordés grâce notamment à l'amitié de Yannick Jacques, Jean Léger et Rémi Houlgatte.

Concernant le premier axe, mes travaux m'ont conduit à m'intéresser à la classification supervisée de protéines, en utilisant les Séparateurs à Vastes Marges ainsi que des méthodes évolutionnaires pour améliorer les performances des classifieurs. Le deuxième axe de recherche m'a offert l'opportunité de me confronter à des problèmes d'analyse statistique de données de puces à ADN. Cette démarche a pu être théorique, en accord avec le cadre de l'Analyse Statistique Implicative, et pragmatique, en réponse à des problématiques soulevées par mes collègues biologistes. J'ai été amené notamment à participer au développement de la panoplie d'outils MADtools de la plate-forme bio-informatique de Nantes. Mes travaux ont ainsi porté sur l'intégration des identifiants de gènes et sur leur analyse par des techniques statistiques, ainsi que sur la méta-analyse de données de puces portant sur des pathologies musculaires.

Mes travaux de recherche sont décrits dans les parties A et B de ce mémoire, reflétant les deux thématiques traitées. La partie C détaille mon *curriculum vitae*, en présentant mon parcours professionnel, mes activités de recherche, d'enseignement et d'administration, ainsi que la liste de mes publications scientifiques.

Organisation générale du mémoire

Introduction. Une tentative de définition de la bio-informatique, en donnant quels brefs rappels historiques et en indiquant ses champs d'application.

p. 1

Partie A. Cette partie traite de la recherche d'homologues distants dans une famille de protéines. Après un bref rappel de notions essentielles de biologie moléculaire, la technique des Séparateurs à Vastes Marges est présentée, suivie d'une adaptation originale en classification de protéines. Une proposition d'un système d'agrégation de classifieurs à base d'essaim grammatical est enfin donnée dans le cadre de cette thématique.

p. 6

Partie B. La deuxième partie expose mes activités de recherche en analyse de données de puces à ADN. Après une présentation des concepts et des techniques biologiques associées à ce type de données, je développe mes travaux relatifs à l'application de la théorie de l'Analyse Statistique Implicative. Le chapitre suivant concerne l'intégration des identifiants de gènes et une approche originale d'analyse de listes d'identifiants pour en extraire des patrons remarquables. Le dernier chapitre indique mon apport sur une méta-analyse de données de pathologies musculaires à partir d'une banque de données publiques consacrée au transcriptome.

p. 35

Conclusion et Perspectives. Réflexion sur les travaux menés et les défis à venir. La méta-analyse à grande échelle et la génomique intégrative.

p. 75

Curriculum Vitae. Présentation du Curriculum Vitae. Parcours professionnel, thèmes de recherche, publications scientifiques, encadrement, responsabilités scientifiques, administratives et d'enseignement.

p. 93

Introduction

La bio-informatique est une discipline relativement récente, le terme ayant été créé dans les années 80. Cette notion englobe l'ensemble des applications de l'informatique aux sciences de la vie, domaine très vaste qui recouvre tous les axes de recherche, allant des applications en robotique aux techniques les plus avancées en intelligence artificielle. Pour la plupart des membres de la communauté scientifique, cette notion semble dans la pratique s'adapter plus particulièrement aux outils informatiques qui permettent de stocker, d'analyser et de visualiser les informations contenues dans les séquences des gènes et des protéines des êtres vivants. L'histoire de la bio-informatique est donc étroitement liée à celle de la biologie moléculaire, l'étude des molécules du vivant.

Il est intéressant de constater que l'essor des connaissances en biologie moléculaire progresse parallèlement à celle de l'informatique (voir quelques jalons historiques en Table 1). En ce qui concerne la biologie moléculaire, un tournant important a été impulsé par la mise au point de techniques de séquençage de l'ADN en 1977 (conjointement par Frederick Sanger d'une part, et par Allan Maxam et Walter Gilbert d'autre part). Il faudra attendre le milieu des années 1980 pour voir apparaître le développement des premiers robots séquenceurs. Dans les mêmes années l'informatique connaîtra de grandes avancées : avènement des micro-ordinateurs personnels et création de langages de programmation évolués (comme le langage C). Les biologistes s'apercevront rapidement du bénéfice qu'ils pourront tirer de tels outils. Il faut en effet se rappeler que les premières recherches en biologie moléculaires ont été menées avec des moyens très limités : en 1962, la résolution de la structure de la myoglobine a demandé à Max Perutz le traçage manuel de plus de deux mille cercles. La visualisation tridimensionnelle des molécules nécessitait par ailleurs la construction fastidieuse d'imposantes structures à base de tiges métalliques qui encombraient les bureaux des chercheurs.

Un des pionniers de la bio-informatique est très certainement Rodger Staden qui a très vite ressenti l'intérêt de développer des programmes pour analyser les séquences. Dès 1977, il propose ainsi des outils informatiques qui servent encore aujourd'hui (un package très utilisé porte son nom). Le résumé d'un de ces articles, intitulé " *Sequence data handling by computer* " [Staden 1977], est très explicite quant à ces motivations :

The speed of the new DNA sequencing techniques has created a need for computer programs to handle the data produced. This paper describes simple programs designed specifically for use by people with little or no computer experience. The programs are for use on small computers and provide facilities for storage, editing and analysis of both DNA and amino acid sequences.

Quelques dates clés en biologie et en informatique

1953
Découverte de la structure de l'ADN
James D. Watson and Francis Crick

1955
Détermination de la structure des protéines
Frédérique Sanger (Prix Nobel 1958)
Découverte de la première séquence protéique : l'insuline

1965
Découverte des mécanismes de la régulation génétique. Jacques Monod, François Jacob et A. Wolf.

1967
Méthode de construction d'arbres phylogénétiques par Fitch et Margoliash

1968
Atlas of protein sequence and structure.
Premier ouvrage contenant 20 séquences de protéines. Margaret Dayhoff

1970
Alignement global de séquences. Algorithme de Needleman et Wunsch.

1977
Méthode de séquençage de l'ADN. Séquençage du premier génome à ADN, le bactériophage phiX174 (5386pb) Frederik Sanger.

1980
Création de la banque EMBL : banque européenne généraliste de séquences nucléiques.

1982
Création de la banque Genbank : banque américaine généraliste de séquences nucléiques.

1987
Apparition de la technologie des puces à ADN.

1990
Programme BLAST (Altschul et al.) : recherche rapide d'alignements locaux dans une banque.

1996
Séquençage du 1er génome eucaryote, *Saccharomyces cerevisiae* (12 Mb, 6000 gènes, 16 chromosomes).

2003
Séquençage complet du génome humain

1951
UNIVAC Premier ordinateur commercialisé. J. Eckert / J. Mauchly

1956
Fortran ((FORmula TRANslation)
Premier langage informatique de haut niveau.

1960
Conception du PDP-1 DEC

1965
Ordinateur IBM/360

1968
Parution de : *The Art of Computer Programming*
Donald Knuth

1969
Conception du système d'exploitation Unix
K. THOMPSON

1970
The traveling-salesman problem and minimum spanning trees. Article de Michael Held and Richard Karp

1971
Premier microprocesseur INTEL

1972
Langage C
D. RITCHIE

1974
CRAY I
Premier super-ordinateur

1977
Invention de l'Apple II micro-ordinateur

1981
PC IBM micro-ordinateur personnel

1986
Induction of decision trees. Machine Learning,
Article de J R Quinlan.

1993
Mosaic (NCSA), premier navigateur INTERNET. 200 sites web disponibles.

1994
Fast algorithms for mining association rules. Article de R Agrawal, R Srikant.

Table 1. *Repères historiques*

La définition exacte du terme bio-informatique constitue une source récurrente de dissensions au sein de la communauté scientifique. Deux approches peuvent être relevées : la première consiste à concevoir la bio-informatique comme un procédé nouveau d'investigation biologique ; la deuxième approche associe cette discipline à un ensemble d'outils mis à la disposition des biologistes pour valider des expériences biologiques.
Les tenants de la première approche considèrent que l'informatique bouleverse fondamentalement la recherche en biologie moléculaire. Elle apporte un nouveau paradigme de recherche défini par la conception de modèles mathématiques sur lesquels peuvent se mener des expériences *in silico*. Ce néologisme a été créé par analogie avec le terme latin *in vivo* pour désigner des simulations numériques dont le but est la découverte de nouvelles lois ou fonctions biologiques, par opposition aux manipulations expérimentales classiques. Un exemple basique de découverte est l'inférence de la fonction d'une protéine à partir de sa séquence primaire, par identification de séquences similaires dont la fonction est connue. Cette conception de la bio-informatique est à rapprocher du terme anglo-saxon *computational biology*, qui accorde une grande part à la théorie et à la modélisation. La deuxième approche consiste à concevoir l'informatique comme un outil d'analyse de données adapté aux besoins des biologistes. Cette discipline conçoit et développe des méthodes et des logiciels pour le stockage et le traitement de données biologiques.

Ce dernier point de vue, partagé par de nombreux biologistes, s'explique aisément par un souci pragmatique de gérer la masse d'informations nouvelles extraite d'expériences à haut débit. La création de bases de données a très certainement été un des premiers objectifs des chercheurs. Les données biologiques progressent en outre à un rythme accru et cette accumulation ne semble pas connaître de limites. Depuis plus d'une décennie on observe une augmentation gigantesque du volume de données disponible. Des bases de données telles que GenBank [Benson 2001] pour les acides nucléiques et SwissProt [Bairoch 2000] pour les protéines ont vu leurs données doubler de taille tous les quinze mois. Début 2010 GenBank renfermait pas moins de 106 533 156 756 bases correspondant à 108 431 692 de séquences ; SwissProt contenait 180 900 945 acides aminés provenant de 186 149 références annotées (la version basée sur de l'annotation automatique (base TrEMBL) contenait 10 158 056 de séquences). L'avènement de nouvelles technologies comme le *Next Generation Sequencing* [Margulies 2005] a considérablement compliqué la donne, puisque certains séquenceurs peuvent engendrer jusqu'à 3 Gigabases par jour [Richter 2009]. Le coût du séquençage devrait baisser et permettre le développement de la génomique individuelle, et donc une véritable révolution dans les pratiques des chercheurs en biologie et des cliniciens. La Table 2 présente quelques-unes des techniques utilisées en biologie moléculaire.

Technologie	**But**
Séquençage ADN	Détermination de l'ordre exact des nucléotides dans un fragment d'ADN, assemblage automatique des fragments, reconstitution d'un génome.
Spectrométrie de masse	Caractérisation des molécules d'après leur masse. Identification de protéines
Puces à ADN	Mesure de l'expression des gènes
Chromatin ImmunoPrecipitation on Chip (ChIP-chip)	Identification des sites de fixation des facteurs de transcription.
ChIP-seq	Technique alternative au ChIP-chip utilisant le séquençage des fragments d'ADN en liaison avec un facteur de transcription.

Table 2. *Eventail de technologies de biologie moléculaire*

L’énorme masse d'informations biologiques doit être gérée par des outils informatiques adaptés. Dans un premier temps, l'effort a porté sur l'organisation de ces données pour définir des normes, des structures, le stockage et l'interrogation des données. Très tôt est apparue la nécessité de concevoir des algorithmes dédiés. Un deuxième axe de recherche s’est donc rapidement développé pour améliorer le traitement de ces données. Fait remarquable, l’un des articles scientifiques les plus cités dans les années 90, toutes disciplines confondues, est celui décrivant l’outil de comparaison de séquences *Basic Local Alignment Search Tool* [Altschul 1990]. BLAST est très certainement l’algorithme le plus utilisé au sein de la communauté des chercheurs en biologie. Une raison de ce succès provient du fait que les séquences de gènes et de protéines sont d’une importance capitale en biologie moléculaire, aussi bien sur les aspects fondamentaux (comme l’étude évolutionnaire par exemple) que pour les conséquences pratiques qui en découlent (notamment en termes de conception de nouveaux médicaments). BLAST permet de comparer deux séquences biologiques et de rechercher des régions similaires entre ces séquences. Cet outil a été conçu pour rechercher rapidement des séquences connues proches de la séquence donnée par l’utilisateur en interrogeant des bases de données spécialisées. On voit à travers cet exemple comment une nouvelle discipline a émergé par intégration de compétences en biologie moléculaire, en algorithmie, en mathématiques et en statistiques.

Les deux approches de la bio-informatique qui ont été mentionnées plus haut ne sont pas à opposer : elles constituent les deux faces d’une même discipline, selon qu’on considère son aspect fondamental (modélisation mathématique) ou selon qu’on s’intéresse aux retombées en termes d’ingénierie (développement d’applications informatiques performantes pour les utilisateurs biologistes).
Les applications les plus courantes de la bio-informatique (voir Table 3) sont l'analyse et la comparaison de séquences, les alignements multiples de séquences, l'identification et la visualisation de structures protéiques, la prédiction de la structure et de la fonction d'une protéine à partir de sa séquence, la comparaison de génomes, l'analyse de l'expression des gènes et des réseaux de régulation. La thématique qui a connu le plus grand développement est sans doute la génomique, à savoir l'étude du fonctionnement biologique à l'échelle du génome. Deux branches sont à distinguer, la génomique structurale et la génomique fonctionnelle. La première branche étudie la structure du génome et du protéome, la deuxième s'intéresse aux fonctions des gènes. Parmi les techniques utilisées figure la transcriptomique, qui est consacrée à l'analyse de l'expression des gènes (*ARNm*).

Sources de données	**Tâches bio-informatiques**
Séquences nucléiques	Identification des introns et des exons Prédiction de gènes
Séquences protéiques	Algorithmes de comparaison de séquences Alignement multiple Découverte de régions conservées Recherche de motifs
Structures macromoléculaires	Prédiction de structures secondaires Prédiction de structures tertiaires Identification d’interactions moléculaires
Génomes	Analyse de la structure du génome Phylogénie Analyse de liaison génétique
Expression des gènes	Corrélation de l’expression des gènes Recherche de gènes différentiellement exprimés
Articles biomédicaux	Annotation automatique des entités biologiques Bases de connaissances biomédicales à partir de texte.
Voies de signalisation métaboliques	Simulation de réseaux

Table 3. *Applications classiques en bio-informatique*

Au sein de ce très large éventail de recherche, je me suis intéressé à l'application de techniques de fouille de données (ou *data mining)* aux entités biologiques. Deux axes ont été développés, le premier a trait à l'identification d'homologues distants au sein d'une famille de protéines, le second repose sur l'analyse des données de puces à ADN.

La recherche d'homologues distants consiste à déterminer si une protéine inconnue appartient à une famille de protéines. Pour des familles protéiques peu homogènes, la similarité est si faible qu'une comparaison de séquences n'est pas pertinente pour assurer l'identification de nouveaux membres. Il est alors nécessaire de concevoir des méthodes de classification capables de discriminer entre familles de protéines. Le premier chapitre présente la problématique biologique sous-jacente, le deuxième un algorithme original de classification et le troisième une approche visant à fusionner plusieurs classifieurs pour améliorer les performances.

La deuxième partie de ce mémoire porte sur l'analyse des données du *transcriptome*. Le quatrième chapitre expose les techniques des puces à ADN ainsi que les mesures associées. Le chapitre suivant est consacré à l'analyse implicative des données d'expression. Je présente dans le sixième chapitre la problématique de la méta-analyse. En conclusion, après avoir rappelé ma contribution dans l'analyse des données de puces à ADN, j'indiquerai les pistes de recherche que je compte aborder dans les prochaines années.

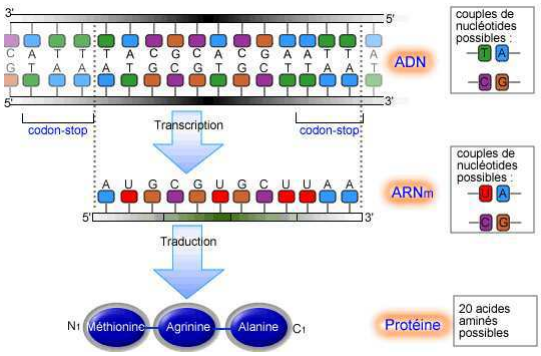
Partie A. Recherche d’homologues distants dans une famille de protéines

Chapitre 1. L’analyse de séquences protéiques

Notre patrimoine génétique est stocké au sein de nos cellules et est défini par l'ensemble de nos gènes ou génome. Plusieurs métaphores ont été données pour caractériser ce phénomène. Une des plus anciennes et des plus fructueuses est celle imaginée par Erwin Schrödinger en 1945 : *It is these chromosomes […] that contain in some kind of code-script the entire pattern of the individual’s future development and of its functioning in the mature state*.

Parallèlement à cette notion de code génétique, de nombreuses analogies ont vu le jour, telles que la recette de cuisine ou le plan d’architecte. La dernière métaphore possède l’intérêt de faire précisément référence à une construction ; les gènes mémorisent un plan pour ériger une protéine, un des matériaux de base les plus essentiels de notre organisme. Rappelons brièvement que les instructions d’un gène sont codées dans un langage très simple, formé par un alphabet de 4 lettres (A, C, G et T). Si l’ADN n’est pas apparu immédiatement comme le support de l’information génétique, c’est que son contenu paraissait trop faible pour renfermer les plans de fabrication du vivant. La découverte de la structure de l'ADN [Watson 1953] a permis de constater que cette macromolécule linéaire forme une séquence composée de quatre bases nucléiques (l'adénosine (A), la cytosine (C), la guanine (G) et la thymine (T)). Les auteurs ont très rapidement saisi l’importance de cette découverte : la richesse informationnelle de cette macromolécule provient de l’extrême variabilité de composition des séquences (le génome humain comporte environ trois milliards de caractères).

L'ADN est stocké dans le noyau de la cellule, qui constitue donc le siège de l'information génétique. La synthèse des protéines ou traduction (transformation de l’information génétique en matériau utile pour la cellule) a lieu dans le cytoplasme grâce à des unités de fabrication appelées ribosomes. Il existe cependant une étape intermédiaire appelée transcription durant laquelle l’ADN est transformé en ARN messager. Comme son nom l’indique, cette molécule sert de support de l’information en passant à travers la membrane nucléaire et en transportant le code nécessaire à la synthèse de la protéine dans le cytoplasme. Ce mécanisme peut être assimilé à la photocopie d’un plan pour éviter de dégrader l’original. La Fig. 1 résume les deux étapes de construction d’une protéine à partir d’un gène.



Auteur : Guillaume Bokiau

Fig. 1. *Les deux étapes menant de l’ADN à la protéine*

L'ordre des lettres dans un gène est donc capital pour sa traduction en protéine. A l’instar des gènes, les protéines sont des macromolécules formées par une séquence de lettres : les acides aminés,

définis par un alphabet de 20 lettres. La Fig. 2 indique le code génétique liant un codon (séquence de trois bases) et un acide aminé. La séquence d'une protéine, ou la succession des acides aminés qui la composent, conditionne sa forme et donc sa fonction dans l'organisme.

1st base

2nd base	U		C		A		G		3rd base
U	UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine	U
	UUC	Phenylalanine	UCC	Serine	UAC	Tyrosine	UGC	Cysteine	C
	UUA	Leucine	UCA	Serine	UAA	Stop	UGA	Stop	A
	UUG	Leucine	UCG	Serine	UAG	Stop	UGG	Tryptophan	G
C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine	U
	CUC	Leucine	CCC	Proline	CAC	Histidine	CGC	Arginine	C
	CUA	Leucine	CCA	Proline	CAA	Glutamine	CGA	Arginine	A
	CUG	Leucine	CCG	Proline	CAG	Glutamine	CGG	Arginine	G
A	AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine	U
	AUC	Isoleucine	ACC	Threonine	AAC	Asparagine	AGC	Serine	C
	AUA	Isoleucine	ACA	Threonine	AAA	Lysine	AGA	Arginine	A
	AUG	Methionine (Start)	ACG	Threonine	AAG	Lysine	AGG	Arginine	G
G	GUU	Valine	GCU	Alanine	GAU	Aspartic Acid	GGU	Glycine	U
	GUC	Valine	GCC	Alanine	GAC	Aspartic Acid	GGC	Glycine	C
	GUA	Valine	GCA	Alanine	GAA	Glutamic Acid	GGA	Glycine	A
	GUG	Valine	GCG	Alanine	GAG	Glutamic Acid	GGG	Glycine	G

Nonpolar, aliphatic Polar, uncharged Aromatic Positively charged Negatively charged

Fig. 2. *Le code génétique.*
Les 20 acides aminés sont indiqués dans la table selon l'ordre des nucléotides dans le codon.

La Fig. 3 montre un classement des acides aminés en fonction de leurs propriétés physico-chimiques, tel qu'il a été établi dans [Taylor 1986]. La mutation d'un acide aminé en un autre est d'autant plus facilité que les deux molécules partagent des propriétés communes.

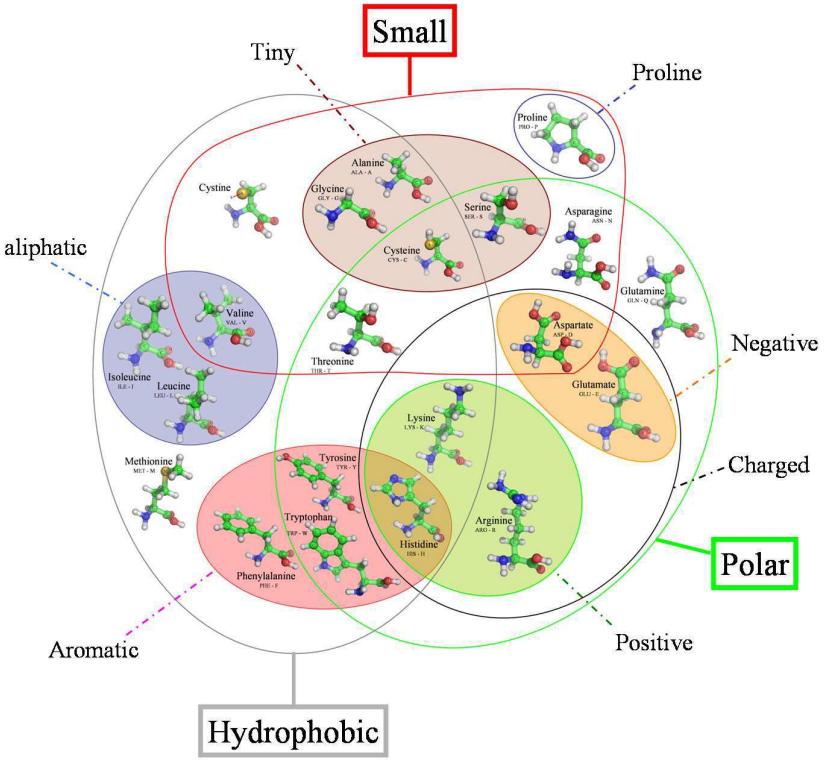


Fig. 3. *Classement des acides aminés selon leurs propriétés physico-chimiques.*

La Fig. 4 indique un exemple de séquence nucléotidique correspondant à l'interleukine IL1 (CDS (*Coding DNA Sequence*) de 816 paires de base), ainsi que sa séquence protéique associée.

```
   1 atggccaaag ttccagacat gtttgaagac ctgaagaact gttacagtga aaatgaagaa
  61 gacagttcct ccattgatca tctgtctctg aatcagaaat ccttctatca tgtaagctat
 121 ggcccactcc atgaaggctg catggatcaa tctgtgtctc tgagtatctc tgaaacctct
 181 aaaacatcca agcttacctt caaggagagc atggtggtag tagcaaccaa cgggaaggtt
 241 ctgaagaaga gacggttgag tttaagccaa tccatcactg atgatgacct ggaggccatc
 301 gccaatgact cagaggaaga aatcatcaag cctaggtcag caccttttag cttcctgagc
 361 aatgtgaaat acaactttat gaggatcatc aaatacgaat tcatcctgaa tgacgccctc
 421 aatcaaagta taattcgagc caatgatcag tacctcacgg ctgctgcatt acataatctg
 481 gatgaagcag tgaaatttga catgggtgct tataagtcat caaaggatga tgctaaaatt
 541 accgtgattc taagaatctc aaaaactcaa ttgtatgtga ctgcccaaga tgaagaccaa
 601 ccagtgctgc tgaaggagat gcctgagata cccaaaacca tcacaggtag tgagaccaac
 661 ctcctcttct tctgggaaac tcacggcact aagaactatt tcacatcagt tgcccatcca
 721 aacttgttta ttgccacaaa gcaagactac tgggtgtgct tggcaggggg gccaccctct
 781 atcactgact ttcagatact ggaaaaccag gcgtag

MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYHVSYGPLHEGCMDQSVSLSISETSKTSKLTFK
ESMVVVATNGKVLKKRRLSLSQSITDDDLEAIANDSEEEIIKPRSAPFSFLSNVKYNFMRIIKYEFILND
ALNQSIIRANDQYLTAAALHNLDEAVKFDMGAYKSSKDDAKITVILRISKTQLYVTAQDEDQPVLLKE
MPEIPKTITGSETNLLFFWETHGTKNYFTSVAHPNLFIATKQDYWVCLAGGPPSITDFQILENQA
```

Fig. 4. *Séquence codante d'un gène et protéine associée.*
La protéine en question est l'interleukine 1 alpha chez l'homme (information extraite de la banque de données GenBank). La séquence du haut correspond au gène et décrit l'ordre des bases nucléiques (la première colonne indique la position dans la séquence ; la présentation en plusieurs lignes et par groupe de 10 lettres est uniquement destinée à faciliter la lecture de la séquence). La séquence du bas correspond à la protéine codée par le gène.

Les propriétés d'une protéine ne sont pas uniquement liées à la structure primaire (enchaînement des acides aminés) mais également à sa conformation dans l'espace (structure tridimensionnelle). La forme d'une protéine détermine dans une large mesure sa fonction, en assurant par exemple la fixation d'un ligand (molécule capable de « reconnaître » une autre molécule). Ce mécanisme aura pour conséquence le déclenchement d'une réponse cellulaire.

La protéine est par ailleurs soumise à des repliements locaux qui induisent la structure tridimensionnelle. Ces repliements prennent certaines formes spécifiques : on distingue les hélices alpha, les brins et feuillets beta, les coudes. La structure en hélice alpha est une structure héliocoïdale périodique, les brins et feuillets beta sont d'autres formes périodiques, les coudes étant des structures apériodiques souvent présentes entre deux structures périodiques. On appelle structure secondaire la séquence de la protéine annotée par ces trois types de repliements. Celle-ci est donc modélisée sous la forme d'un enchaînement de symboles défini par un alphabet de trois lettres : hélice (H), feuillet (E) et coude (C). De nombreuses méthodes bio-informatiques ont vu le jour pour prédire la structure secondaire à partir de la structure primaire. On sait en effet que certains groupes d'acides aminés favorisent la formation d'un type particulier de structure secondaire.

Les méthodes de prédiction de structures secondaires sont nombreuses [Rost 2001]. Outre des méthodes statistiques ou des méthodes basées sur les Plus Proches Voisins, ont souvent été utilisés des Modèles de Markov cachés : les chaînes de Markov modélisent les trois types de structure secondaire par apprentissage sur des séquences dont la structure secondaire est connue. Un score est ensuite associé à chacun des modèles et la structure retenue est celle ayant obtenu le meilleur score. Il existe également des méthodes d'apprentissage utilisant des réseaux de neurones. Le principe de cette technique est analogue à celui employé pour entraîner les Modèles de Markov cachés. D'autres classifieurs ont été employés plus récemment, dont des Séparateurs à Vastes Marges. Les

logiciels de dernière génération sont basés sur des fusions de classifieurs pour accroître les performances de prédiction.

La prédiction de la structure tertiaire (structure tridimensionnelle) est un problème encore plus complexe et loin d'être résolu. Des méthodes ont été proposées, essentiellement basées sur l'homologie avec des protéines dont la séquence tertiaire est connue (notamment par cristallographie par rayon X). La Fig. 5 présente un exemple de conformation d'une protéine dans l'espace, en relation avec sa structure secondaire.

Il existe enfin un dernier niveau : la structure quaternaire liée à l'assemblage de plusieurs chaînes polypeptidiques, qui sont des sous-unités structurales de la protéine finale.

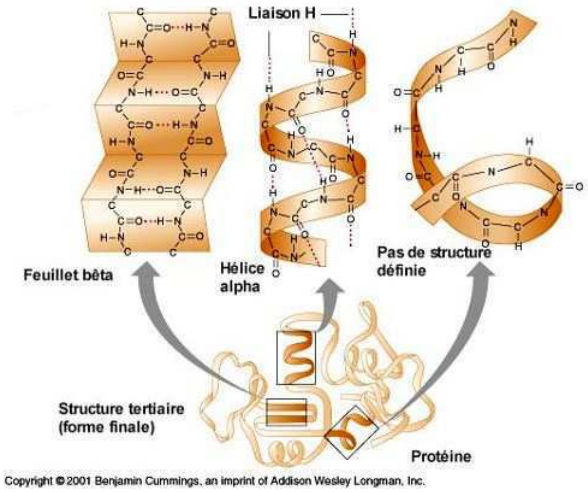


Fig. 5. *Les structures secondaires et tertiaires de la protéine*

La découverte de la fonction d'une protéine est un axe de recherche majeur en génomique, puisque les processus biologiques sont activés par ces molécules. Par exemple, l'hémoglobine est une protéine assurant le transport de l'oxygène dans le sang. La connaissance de la séquence d'une protéine est aussi importante pour déterminer des anomalies pathogènes. La modification d'un acide aminé peut dans certains cas avoir des conséquences néfastes. Ainsi, la drépanocytose est une maladie caractérisée par l'altération de l'hémoglobine. Cette mutation entraîne de nombreux symptômes : anémies, infections et risques d'accidents vasculaires cérébraux. Cette maladie qui touche 50 millions d'individus dans le monde est causée par la modification d'un seul acide aminé.

Les protéines se regroupent en familles : leurs membres sont des protéines partageant généralement un ancêtre commun, chacun des descendants ayant évolué au fil de temps. On parle d'homologie de séquences pour caractériser ce phénomène. Un indice essentiel pour conjecturer une homologie consiste à calculer une similarité de séquences. Il faut cependant éviter d'assimiler les deux notions. L'homologie exprime en effet une réalité biologique, à savoir l'appartenance de deux séquences à une famille commune. La similarité n'est qu'une similitude de séquences établi par un algorithme. Comme il se peut qu'un motif commun apparaisse entre deux séquences prises au hasard, le terme d'homologie n'est pas approprié pour définir le résultat d'un algorithme de comparaison de séquences : on préférera parler de pourcentage de similarité. On admet cependant généralement qu'un pourcentage de similarité important (supérieur à 30%) est un fait significatif permettant de présumer l'existence d'une homologie entre deux séquences.

La base SCOP [Murzin 1995] de protéines regroupe l'ensemble des protéines connues. Les méthodes automatiques de comparaison de séquences et de structures ne permettent pas toujours d'identifier correctement les relations évolutionnaires entre les protéines. La base SCOP a donc a été construite manuellement pour catégoriser les protéines, à travers de nombreux outils, dont des outils de visualisation pour faciliter la tâche des experts.

La base de données SCOP repose sur une organisation hiérarchique en trois niveaux :

- Niveau 1 : la famille. Les protéines regroupées à ce niveau sont certainement liées entre elles par un lien évolutionnaire. Généralement, les séquences sont très similaires (30% de similarité), mais les experts ont également classé des familles relativement hétérogènes. Les globines forment par exemple une famille ayant une similarité faible (15% de similarité), mais les similarités de structure et de fonction biologique sont assez fortes pour qu'on puisse regrouper ces protéines au sein d'une même famille.
- Niveau 2 : la superfamille. Les protéines d'une même superfamille possèdent un lien de similarité plus faible, mais des similitudes suffisantes en termes de structure et de fonction pour suggérer l'existence probable d'un lien évolutionnaire.
- Niveau 3 : la similarité structurelle. Ce classement regroupe les protéines en catégories selon le type de repliement. Les protéines partagent des structures secondaires communes dans le même arrangement et avec les mêmes connections topologiques.

La problématique que nous avons abordée est liée à la classification de protéines. Ce travail a été mené en collaboration avec Yannick Jacques, responsable de l'équipe *Cytokines et récepteurs en immuno-cancérologie* de l'unité INSERM U892. Cette équipe s'intéresse à une superfamille de protéines appelées cytokines. Les cytokines sont une classe de protéines ayant un rôle dans la communication intercellulaire et plus particulièrement dans la réaction immunitaire. Les gènes de cette superfamille sont souvent pleïotropiques (une mutation dans un seul gène induit plusieurs phénotypes mutants) et redondants, avec des effets agonistes ou antagonistes. C'est une superfamille qui possède une faible similarité de séquence au niveau nucléique et protéique. Cette faible similarité explique la difficulté à trouver les homologues distants dans cette famille. Nous nous intéressons plus spécifiquement aux sous-familles IL-6, IL-2 et IL-10/INF, constituées principalement de structures quatre hélices alpha.

Le chapitre suivant décrit une méthode originale d'identification d'homologues distants basée sur les Séparateurs à Vastes Marges. Le chapitre 3 présente un système complet de classification de protéines à partir de plusieurs classifieurs.

Chapitre 2. Identification d'homologues distants dans une famille de protéines

Nous présentons dans ce chapitre une méthode discriminative pour la classification de protéines dans le cas des familles présentant une homologie faible entre membres. Cette restriction s'explique par le fait que des similarités fortes entre séquences permettent de recourir à une procédure simplifiée. Il suffit alors d'extraire des motifs plus ou moins conservés. Ces motifs une fois isolés, il devient facile de rechercher la présence de ces motifs sous une forme exacte ou dégénérée dans des séquences candidates, à l'aide d'un logiciel comme ScanProsite [de Castro 2006]. Les familles distantes sont plus difficiles à caractériser et nécessitent des méthodes appropriées. Celle que nous proposons est basée sur une modélisation de la famille par un ensemble $\mathcal{M}$ de motifs de faibles tailles. Chacun de ces motifs considérés de façon indépendante est trop peu discriminant : un motif individuel peut se retrouver dans une séquence biologique par le seul jeu du hasard. La présence de nombreux motifs peut cependant être l'indice d'une forme de similarité plus subtile qui ne peut être obtenue par des scores classiques de comparaison de séquences. Comme nous le verrons plus loin, l'originalité de notre jeu de motifs provient de l'utilisation des propriétés physico-chimiques des acides aminés.

L'ensemble $\mathcal{M}$ définit un espace de caractéristiques sur lequel peuvent être projetées les séquences protéiques. Celles-ci sont transformées en un vecteur de taille égale au cardinal de $\mathcal{M}$, chaque composante de ce vecteur représentant l'appartenance ou non du motif au sein de la séquence. Nous utilisons ensuite la méthode des Séparateurs à Vastes Marges (SVM) pour discriminer la famille cible contre les protéines non apparentées. L'ensemble $\mathcal{M}$ des motifs est préalablement obtenu par apprentissage basé sur des séquences connues extraites de la base de données SCOP [Murzin 1995].

Les sections suivantes présentent un état de l'art des techniques de classification de protéines distantes, le principe de construction de l'ensemble des motifs, l'algorithme d'extraction de ces motifs et enfin les performances de classification de notre classifieur.

2.1. La classification de protéines par Séparateurs à Vastes Marges

Les Séparateurs à Vastes Marges [Vapnik 2000] font partie des techniques de classification supervisée les plus performantes. Elles ont été abondamment utilisées en analyse de données et son usage en bio-informatique est très fréquent. On peut citer à titre d'exemple la classification de tissus cancéreux à partir de données de puces à ADN [Furey 2000], la reconnaissance de repliement de protéines [Ding 2001] ou la prédiction de structure secondaire des protéines [Hua 2001]. Les SVMs demeurent la méthode la plus aboutie pour identifier des protéines distantes [Jaakkola 2000], [Leslie 2002], [Liao 2003], [Leslie 2004], les performances obtenues surpassant les méthodes précédentes reposant sur des modèles de Markov cachés [Karplus 1998].

La classification supervisée repose sur un ensemble de données d'apprentissage et de test qui constitue autant d'instances pour lesquelles la classe d'appartenance est connue. Le problème de classification consiste à identifier une valeur cible (la classe) à partir d'un jeu d'attributs (les caractéristiques associées aux individus). Nous nous bornons ici à rappeler le principe du SVM en classification binaire (appartenance ou non à la famille de protéines cibles).

La classification binaire peut se modéliser sous la forme suivante : soit X un espace d'entrée décrivant des instances et Y le domaine des classes d'appartenance, à savoir $Y= \{-1,+1\}$. Au lieu de chercher directement à construire une fonction d'étiquetage $E : X \rightarrow Y$, on définit une fonction à valeur dans **R** :

$$f: X \rightarrow \mathbf{R} \qquad (1)$$

La classe se détermine alors par la fonction $h = \mathrm{sgn}(f)$ (2)

La Fig. 6 représente le principe fondamental des SVMs, à savoir la séparation de deux classes par un hyperplan maximisant la marge entre les deux populations. La définition de l'hyperplan et des vecteurs-supports est un problème d'optimisation relativement complexe (voir [Vapnik 2000] pour une présentation formelle des SVMs). Nous nous bornerons à aborder ici l'idée intuitive sous-jacente, à savoir que la discrimination est d'autant meilleure que la marge est importante. Les individus les plus proches constituent les vecteurs supports : ce sont eux qui déterminent la séparation maximale. La fonction h est donc associée à la position des individus par rapport à la frontière de décision.

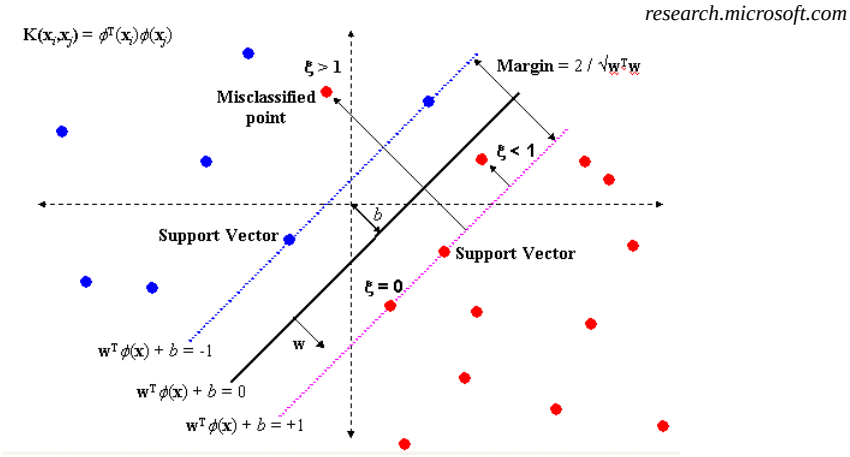


Fig. 6. *Principe géométrique du SVM.*
Cette image représente deux classes (dont les instances sont représentées en bleu et en rouge) ainsi que la frontière optimale (hyperplan) séparant celles-ci.

Une limite évidente d'une telle méthode est qu'il n'est pas toujours possible de séparer linéairement deux classes, comme le montre l'exemple de la Fig. 7.

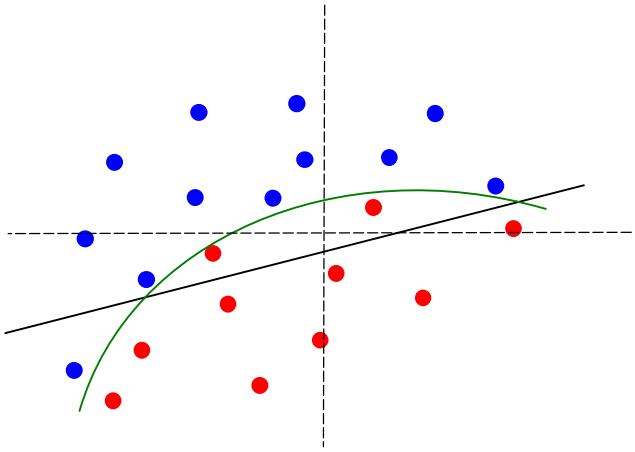


Fig. 7. *Séparabilité des classes par un hyperplan*
Sur cet exemple, il n'existe pas de frontière de décision linéaire séparant exactement les deux classes. La courbe verte offre une meilleure séparation entre les deux populations.

Pour dépasser cette limitation, les SVMs opèrent une transformation sur les individus en les projetant dans un espace dénommé espace des caractéristiques :

$$\Phi : X \rightarrow \mathcal{F} \qquad (3)$$

Cet espace nouveau doit être de dimension supérieure à l'espace d'entrée, voire de dimension infinie. Il serait très difficile d'implémenter de tels espaces, mais une astuce permet de s'affranchir de cette difficulté. En effet, la définition de l'hyperplan et la position des individus par rapport à lui sont uniquement déterminés par des produits scalaires. Dans la pratique, on remplace donc le produit scalaire dans l'espace des caractéristiques par une fonction dénommée fonction noyau :

$$\Phi(x).\Phi(x') = K(x, x') \qquad (4)$$

Certaines propriétés sont requises pour définir une fonction noyau (conditions de Mercer ; voir [Vapnik 2000] pour plus de détails), qu'il n'est pas toujours aisé de démontrer. Plusieurs fonctions ont été proposées, qui ont été retenues dans la pratique : nous avons repris la plus commune d'entre elles, la fonction à base radiale:

$$K(x,x') = \exp(-\| x - x' \|^2/\sigma^2) \qquad (5)$$

Les SVMs sont basés sur des vecteurs numériques de taille fixe. Les séquences protéiques étant des séquences de caractères (alphabet composé des 20 acides aminés) de taille variable, il est nécessaire d'opérer une transformation de ces séquences en vecteurs. Cette stratégie, plus aisée à entreprendre que la recherche d'une fonction noyau dédiée, a été celle qui a été adaptée dans les algorithmes les plus usités en matière d'identification d'homologues. On peut classer ces méthodes en deux classes : les méthodes procédant à une transformation directe des séquences et les méthodes s'appuyant sur des comparaisons de séquences à des jeux d'apprentissage.

Les méthodes de transformation directe ont l'avantage de la simplicité : elles ne prennent en compte que l'information de la séquence candidate. La première méthode de vectorisation proposée a été celle de Spectrum Kernel [Leslie 2002]. Cette technique projette la séquence dans l'espace des mots de taille *k*, *k* étant un paramètre fixé par l'utilisateur. Pour un alphabet {A,B,C}, l'ensemble des mots de taille *k*=3 est { AAA, AAB, AAC, ABA, …, CCC}. Dans notre cas, l'espace d'entrée possède une dimension de 20^k. Pour des grandes valeurs de *k*, l'espace d'entrée peut être prohibitif, mais il s'avère que des valeurs faibles (*k* de l'ordre de 3 ou 4) sont suffisantes pour discriminer les familles de protéines, d'après les résultats des auteurs ainsi que nos propres tests. Le vecteur résultat

est un vecteur de booléens précisant pour chaque composante l'existence du mot associé dans la séquence. L'algorithme de vectorisation repose sur une fenêtre glissante de taille *k* ; on ne prend pas en compte les occurrences multiples ni les positions relatives des mots au sein de la séquence.

Une extension de la technique précédente a été proposée dans l'algorithme Mismatch [Leslie 2004]. Cette méthode considère des similitudes partielles entre une sous-séquence et les mots associés. En autorisant par exemple une erreur (mésappariement ou *mismatch*), la sous-séquence ABC sera associé aux composantes suivantes du vecteur : ABA, ABB, AAC, ACC, BBC, CBC. Les auteurs montrent que ce nouvel algorithme améliore sensiblement les résultats par rapport à la version de base.

La deuxième famille de méthodes de vectorisation est basée sur une comparaison entre une séquence candidate *c* et un jeu de références *R* (un ensemble de séquences dont on connaît la classe). Cette catégorie d'algorithmes suppose la connaissance d'une méthode de comparaison de séquences. Celle-ci permet de définir un score entre la séquence *c* et une séquence s_i appartenant à *R*. L'espace d'entrée est donc formé par *n* composantes, *n* étant la dimension de *R*. La $i^{ème}$ composante du vecteur résultat représente le score entre *c* et s_i.

La première méthode reposant sur ce principe a été Pairwise [Liao 2003] : cette méthode utilise le score de Smith & Watermann [Smith 1981], une technique très connue de comparaison locale de séquences biologiques. La complexité de cet algorithme étant quadratique, son utilisation est plus coûteuse lorsqu'il s'agit de classer un grand nombre de séquences candidates. Une autre mesure de similarité a été considérée dans la même étude : le score donné par un autre algorithme fort utilisé pour la recherche dans les bases de données, à savoir BLAST [Altschul 1990]. La technique de vectorisation s'appuyant sur BLAST a été dénommée Pairwise Blast [Liao 2003]. LA Kernel [Saigo 2004] est une sophistication de la méthode Pairwise : elle reprend le score de Smith & Watermann en considérant tous les alignements locaux possibles en lieu et place du seul optimum comme dans la méthode initiale. Cette méthode semble être la plus performante parmi les techniques que nous avons présentées, mais également la plus coûteuse en temps d'exécution.

2.2. Mesure de performances pour la classification de protéines

La mesure de performances la plus usitée en matière de classification supervisée repose sur le taux d'erreurs obtenu à partir d'un jeu de test. La validation croisée est une technique d'estimation de la précision d'un modèle prédictif. Cette méthode est basée sur le partitionnement des données en deux sous-ensembles complémentaires. Le modèle est appris sur un sous-ensemble dénommé ensemble d'apprentissage et la validation est l'analyse des prédictions opérées sur son complémentaire, dénommé ensemble de test. Pour réduire la variabilité de la mesure, on intervertit jeu d'apprentissage et jeu de tests, et on applique différents partitionnements. Le résultat final est obtenu par la moyenne des différentes validations.

La technique du *leave-one-out* est un cas particulier de validation croisée pour laquelle le jeu de test est réduit à un seul individu. Il est recommandé dans les situations comportant peu de données expérimentales. Comme la famille des cytokines est réduite, nous avons retenu cette mesure de performances dans ce chapitre.

Le chapitre suivant utilise une mesure plus fine de performances : l'aire sous la courbe ROC (AUC ou *Area Under ROC (Receiver Operating Characteristic) curve*). La mesure du taux d'erreurs n'est en effet pas toujours satisfaisante pour discriminer entre plusieurs classifieurs comme le démontre la Fig. 8 : des classements de qualités très différentes se voient notés avec un score identique.

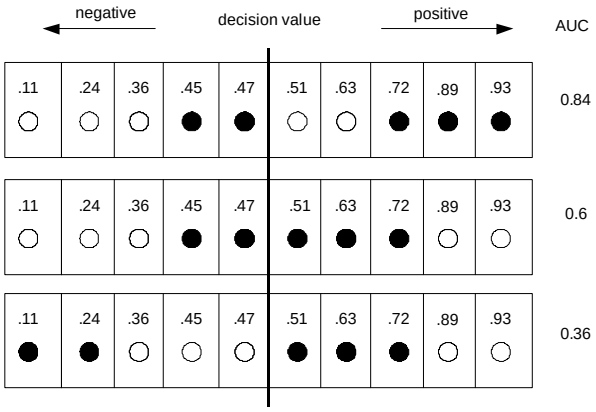


Fig. 8. *Exemples de score AUC.*
La figure représente trois jeux de tests comprenant 10 individus (5 exemples positifs représentés par le symbole • et 5 exemples négatifs représentés par le symbole ◦). Les individus sont ordonnés selon l'ordre croissant de leur probabilité d'appartenance, délivré par un classifieur. Ces trois jeux se voient attribués la même frontière de décision : si la probabilité est supérieure à 0.5, alors on attribue à l'individu considéré la classe positive ; dans le cas contraire, l'individu est étiqueté dans la classe négative. On constate que les trois classements possèdent le même taux d'erreurs de 40%, alors même que les classements sont de qualités distinctes, comme le montrent les valeurs AUC.

Comme son nom l'indique, l'aire sous la courbe ROC est basée sur un modèle graphique dans lequel les abscisses représentent les faux-positifs et les ordonnées les vrais-positifs (voir Fig. 9). Un point de la courbe dénote un résultat de classification pour un seuil de discrimination donné (par exemple dans la Fig. 8, le seuil retenu correspond à une probabilité de 0.5). Par conséquent, l'aire sous la courbe est un bon estimateur de la qualité globale du classifieur, puisqu'il intègre l'ensemble des seuils de discrimination possible. De plus, on montre [Yan 2003] que cette mesure peut être interprétée comme la probabilité qu'un exemple positif possède un score de classification supérieur à celui d'un exemple négatif (les deux exemples ayant été tirés au hasard parmi l'ensemble du jeu de test). Ainsi, un AUC de 0.5 correspond à la performance d'un classifieur aléatoire alors qu'un AUC de 1 est équivalent à un classifieur parfait. Un AUC de 0 dénote enfin un classifieur qui inverse systématiquement la classe à prédire.

Il est nécessaire pour utiliser le score AUC de pouvoir établir un classement des individus. Les SVMs proposent généralement une mesure de distance entre l'individu et la frontière de décision qui peut servir à classer les exemples. Certaines implémentations proposent également une mesure s'apparentant à une probabilité d'appartenance à la classe positive.

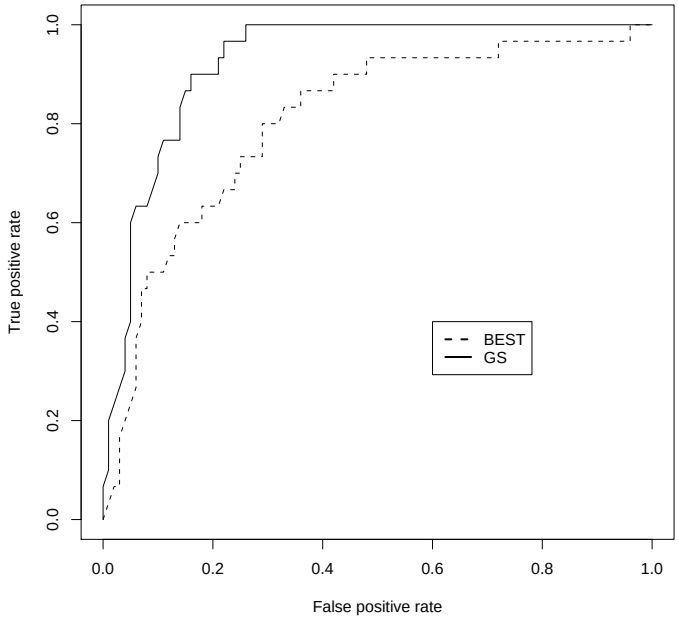


Fig. 9. *Courbes ROC.*

Un classifieur binaire détermine une estimation de l'appartenance d'un individu à la classe positive et utilise un seuil de discrimination pour séparer les individus en deux classes. Cette valeur de seuil constitue une frontière de décision au delà de laquelle un individu sera considéré comme positif. Parmi ceux-ci, certains seront des vrai-positifs (classification correcte) et d'autres des faux-positifs (ce sont en fait des individus négatifs). Chaque point de la courbe ROC définit un point de coupe (seuil de discrimination) pour lequel on a calculé la proportion de vrai-positifs et de faux-positifs. Un bon classifieur correspond à une courbe qui possède une pente très forte pour les faibles valeurs de x et qui épouse assez tôt la droite constante y=1. Sur cet exemple, le classifieur GS est meilleur que le classifieur dénommé BEST (voir le chapitre suivant pour une description de ces deux classifieurs).

2.3. Classification de protéines à partir d'une hiérarchie de motifs

Nous avons proposé un classifieur SVM basé sur la transformation directe d'une séquence primaire de protéines en un vecteur numérique. Contrairement à l'algorithme Spectrum Kernel [Leslie 2002] qui analyse tous les mots de taille *k* possibles, nous avons défini un dictionnaire des mots de taille *k* les plus discriminants pour notre famille d'intérêt (on appellera motifs de telles courtes séquences).

L'alphabet Ω des acides aminés est composé des caractères suivants :

Ω = { A,C,D,E,F,G,H,I,K, L,M,N, P,Q,R, S, T, V,W, Y} (6)

Dans Mismatch [Leslie 2004] sont autorisés des mésappariements pour éviter le sur-apprentissage. Cependant, toutes les variations n'ont pas la même probabilité d'apparaître au sein des homologues. Il est connu que les mutations les plus fréquentes interviennent dans des ensembles d'acides aminés partageant des propriétés physico-chimiques communes. Nous avons donc repris la classification de Taylor présentée sur la Table 4 pour étendre l'alphabet composant les motifs. Cet alphabet est formé par des sous-ensembles de Ω (ou classes de Ω pour reprendre la désignation usuelle): à une position donnée dans le motif, on associe un sous-ensemble qui définit les acides aminés acceptés. Par exemple, un motif comportant la classe α= {I,L,V} acceptera les acides aminés aliphatiques, à savoir I, L ou V. L'alphabet étendu Γ(Ω) que nous avons défini est le suivant :

$$\Gamma(\Omega) = S(\Omega) \cup C(\Omega) \cup \Omega \quad (7)$$

Où $S(\Omega)$ désigne l'ensemble des singletons de Ω : {{A}, {C},…,{Y}} et $C(\Omega)$ l'ensemble des classes de Taylor :

$$C(\Omega) = \{\alpha, \beta, \gamma, \delta, \varepsilon, \varsigma, \eta, \theta\} \quad (8)$$

Symbole	Classe	Membres
α	aliphatic	ILV
β	aromatic	FHWY
γ	non-polar	ACFGHIKLMVWY
δ	charged	DEHKR
ε	polar	CDEHKNQRSTWY
ζ	positive	HKR
η	small	ACDGNPSTV
θ	tiny	ACGST

Table 4. *Classification de Taylor.*
Les acides aminés sont regroupés selon leurs propriétés physico-chimiques. Les symboles grecs associés sont une dénomination personnelle.

Un motif est défini par une séquence d'éléments de Γ(Ω). Par souci de concision, on désignera abusivement les singletons par l'acide aminé associé. Ainsi au motif suivant AαY correspondra les mots suivants : AIY, ALY et AVY. Plus précisément, une occurrence d'un motif $m = \langle m_1, m_2, \ldots, m_k \rangle$ est une sous-séquence $\langle s_i, s_{i+1}, \ldots, s_{i+k-1} \rangle$ d'une séquence *s* telle que $s_{i+j} \in m_j \ \forall\ j, 0 \le j \le k-1$.

On dira d'une séquence comportant au moins une occurrence de *m* qu'elle vérifie ce motif. Par

exemple, la séquence MH vérifie au total 21 motifs, dont les motifs suivants : MH, M β, γδ et ΩΩ. Inversement, il n'existe qu'une seule sous-séquence possible contenant le motif MH tandis que le motif ΩΩ sera vérifié par n'importe quelle sous-séquence de taille 2. Les motifs étant plus ou moins discriminants, il est important d'évaluer leur spécificité, à savoir la probabilité de rencontrer tel motif dans une séquence n'appartenant pas à la famille d'intérêt (nous avons considéré l'ensemble de la base SCOP pour les estimations, à l'exception de la superfamille des cytokines). Pour simplifier cette estimation, nous supposerons que toutes les positions d'une séquence protéique sont issues d'un tirage aléatoire indépendant. Même si cette hypothèse est abusive, elle suffit pour pondérer efficacement les motifs. Nous avons défini la fonction de coût *c*(*m*) suivante :

$$c(m) = \prod_{i=1}^{k} f(m_i) \qquad (9)$$

où $f(m_i)$ est la fréquence de la classe m_i dans notre base d'apprentissage.

Dans la pratique, nous avons observé une bonne corrélation entre cette fonction de coût et les fréquences d'apparition effective des motifs dans la base d'apprentissage.

La spécificité d'un motif *m* sera définie par la fonction sp(*m*)=- log(c(*m*)).

Le *support* d'un motif au sein d'un ensemble Ψ de séquences est défini par le nombre de séquences vérifiant ce motif. Il est important de noter que spécificité et support sont des notions antinomiques : un motif ayant un large support aura une faible spécificité et inversement.

Les motifs sont organisés hiérarchiquement suivant une relation de généralisation. Soit m^1 et m^2 deux motifs. On établit la relation $m^1 \prec m^2$ si pour tout $i \in [1,k]$ $m_i^1 \subseteq m_i^2$

La définition de sp(*m*) implique que sp(m_1)≥sp(m_2) pour des motifs vérifiant $m^1 \prec m^2$. L'ensemble partiellement ordonné $(\Gamma(\Omega), \subseteq)$ forme un sup demi-treillis : toute paire (*x*,*y*) de Γ(Ω)x Γ(Ω) possède une borne supérieure notée sup(*x*,*y*). Par extension on appellera $m^{1,2} = \sup(m^1, m^2)$ la borne supérieure des deux motifs m^1 et m^2 vérifiant $m_i^{1,2} = \sup(m_i^1, m_i^2)$ pour tout $i \in [1,k]$. Le motif $m^{1,2}$ représente le motif le plus spécifique qui généralise m^1 et m^2 : toute séquence vérifiant m^1 et m^2 vérifie $m^{1,2}$. La Table 6 montre des exemples de telles bornes supérieures.

L'algorithme de découverte de motifs fonctionne en deux étapes :

1. Extraction de motifs germes
2. Génération de motifs hiérarchiques

La première étape consiste à identifier des motifs germes à partir d'un ensemble Ψ de séquences. Un motif germe est un motif composé uniquement de singletons (alphabet des acides aminés à l'exclusion des classes physico-chimiques). La recherche de ces motifs réutilise l'algorithme Spectrum Kernel en sélectionnant tous les mots de taille *k* ayant un support suffisant.

La deuxième étape est une construction de motifs par agglomération, selon le principe de la classification hiérarchique ascendante. L'algorithme de génération de motifs est présenté dans la Table 5. La Table 6 présente les motifs hiérarchiques de taille *k*=4 extraits à partir des cinq séquences suivantes :

- HIWY
- HIDY
- KLTY
- HVSG
- DARG

Algorithme *DécouverteMotifs*
Entrées

M, l'ensemble des motifs germes issus de l'étape 1
supMin, le seuil de support minimal
speMin, le seuil de spécificité minimale

Sortie

E, l'ensemble de motifs hiérarchiques

début

$E=\{m \in M \mid \text{support}(m) \geq \text{supMin} \wedge \text{sp}(m) \geq \text{speMin}\}$

Répéter

Soit m^1 et m^2 les motifs de *M* vérifiant :

1. $m^{1,2} = \sup(m^1, m^2)$
2. $\text{sp}(m^{1,2}) \geq \text{sp}(m^{i,j})$ pour tout m^i et m^j dans *M*

$M \leftarrow M - \{m^1, m^2\}$

$M \leftarrow M \cup \{m^{1,2}\}$

Si $\text{support}(m^{1,2}) \geq \text{supMin}$ et $\text{sp}(m_{1,2}) \geq \text{speMin}$ alors

$E \leftarrow E \cup \{m^{1,2}\}$

Jusqu'à cardinal(*M*)=1 ou $\text{sp}(m_{1,2}) < \text{speMin}$

fin.

Table 5. *Algorithme de découverte de motifs hiérarchiques.*

motif	contenu	support	spécificité
m1	HIWY	1	14,25
m2	HIDY	1	12,83
m3	KLTY	1	11,44
m4	HVSG	1	11,79
m5	DARG	1	10,96
$m6=m^{1,2}$	KIεY	2	10,64
$m7=m^{3,6}$	ζαεY	3	7,55
$m8=m^{4,5}$	δηεG	2	5,26
$m9=m^{7,8}$	δΩεγ	5	2,56

Table 6. *Motifs extraits par l'algorithme DécouverteMotifs*
Ces motifs ont été découverts à partir de l'ensemble des séquences {HIWY, HIDY, KLTY, HVSG, DARG}.

Notre jeu de données comprend 45 exemples et 45 contre-exemples. Les exemples sont les interleukines, qui sont des membres de la superfamille des cytokines, et les contre-exemples sont extraits de la base SCOP de sorte à respecter une représentation équilibrée des superfamilles de

SCOP. La taille des motifs a été fixée à 8. Cette valeur a été retenue en raison de la structure héliocoïdale des interleukines (pas d'hélice de l'ordre de 4 acides aminés). Comme le classifieur utilisé est un Séparateur à Vastes Marges, il est nécessaire de vectoriser les séquences. De manière analogue à l'algorithme Spectrum Kernel, l'espace d'entrée est celui des motifs hiérarchiques ; chaque composante correspond à un motif et la valeur booléenne associée indique si la séquence en question vérifie le motif.

La Table 7 présente les résultats de validation croisée obtenus sur ce jeu de données. La méthode KNN est la méthode des plus proches voisins pour laquelle la similarité de séquences est donnée par le produit cartésien des vecteurs booléens. Elle est donnée à titre de comparaison pour évaluer le gain apporté par les SVMs (de l'ordre de 5% de taux d'erreurs en moins). La méthode SKSVM (k=4) est la méthode Spectrum Kernel : c'est la méthode SVM la moins performante parmi celles testées. La méthode MKSVM représente l'algorithme Mismatch (k=4, nombre de mésappariements autorisés=1) qui améliore très légèrement les performances de SKSVM.

Nous indiquons trois résultats de notre méthode SVMotifs, utilisant les trois seuils de spécificité minimale suivants : 13, 14 et 15. Il semblerait que la valeur médiane soit optimale pour notre jeu de données : en deçà de cette valeur, les motifs seraient moins représentatifs et au delà, ces derniers seraient trop spécifiques à certaines cytokines.

classifieur	taux d'erreurs	TP	FN	TN	FP
KNN	18.9	88.9	11.1	73.3	26.7
SKSVM	13.3	84.4	15.6	88.9	11.1
MKSVM	12.0	82.7	17.3	93.3	16.7
SVMotifs13	2.2	95.6	4.4	100	0
SVMotifs14	0	100	0	100	0
SVMotifs15	5.5	88.9	11.1	100	0

Table 7. *Résultats de la validation croisée (leave-one-out).*
Les colonnes TP, FN, TN et FP représentent respectivement les pourcentages de vrais-positifs, de faux-négatifs, de vrais-négatifs et de faux-positifs. Tous les classifieurs SVMs utilisent la fonction noyau à base radiale.

2.4. Conclusion

La recherche d'homologues distants est une problématique ardue en termes de classification supervisée, puisqu'elle implique une optimisation du classement des séquences candidates. L'aire sous la courbe ROC est un bon indicateur de la qualité d'un classement. Des méthodes efficaces en temps d'exécution ont été proposées. Elles définissent des mots (sous-séquence de taille fixe, avec possibilité de mésappariement), permettant ainsi la comparaison de séquences par leur composition en mots élémentaires. Cependant, cette famille de méthodes n'exploite pas assez les connaissances

des propriétés physico-chimiques des acides aminés et ne prend pas en compte les régions conservées des protéines apparentées.

C'est la raison qui nous a amené à proposer une technique de vectorisation utilisant le concept de motif hiérarchique. La technique de classification basée sur de tels motifs présente certains avantages par rapport aux méthodes basées sur des mesures de similarité. Elle est moins coûteuse en temps d'exécution que les méthodes utilisant des jeux de références ; la vérification de la présence d'un motif dans une séquence peut se faire en un temps linéaire alors que les mesures de comparaison de séquences sont en complexité quadratique et difficiles à mettre en œuvre dans la pratique sur de grands jeux de données. Cette méthode s'avère plus efficace que les méthodes analysant la composition d'une séquence en mots, que ceux-ci soient dégénérés ou non.

Ma contribution en matière de classification de protéines a été de concevoir un espace original de représentation des séquences. Contrairement aux techniques de SVMs consistant à utiliser un espace de mots indifférenciés, j'ai proposé une étape préliminaire de découverte de mots sur-représentés dans la famille de protéines d'intérêt. L'algorithme de découverte de ces motifs repose sur des propriétés physico-chimiques générales, ce qui évite le sur-apprentissage. Les très bonnes performances de cette méthode par rapport aux autres SVMs opérant sur des mots montrent que cette approche est pertinente.

Chapitre 3. Agrégation de classifieurs pour la recherche d'homologues

Comme il a été vu dans le chapitre précédent, l'identification d'homologues distants a donné lieu au développement de diverses méthodes. Il est donc possible de confronter les résultats de ces algorithmes pour obtenir de meilleures performances. Cette approche répond à une intuition commune : il est courant de demander l'avis de plusieurs experts dans des situations importantes. Il se peut cependant qu'un avis peu judicieux entrave la qualité des résultats plutôt qu'elle n'améliore celle-ci.

En analyse de données, la fusion de classifieurs est une pratique fortement répandue. Certains auteurs appliquent le vote à la majorité [Kuncheva 2003], d'autres utilisent des techniques de *boosting* [Kearns 1988] ou encore font évoluer des ensembles de prédicteurs par programmation génétique [Brameier 2001].

Nous proposons une approche basée sur la recherche d'un opérateur d'agrégation. Nous employons ce terme dans son sens le plus large pour désigner un opérateur combinant les résultats de plusieurs classifieurs pour former un score unique. Un avantage de cette méthode est de fournir un algorithme explicite de classement. Il peut donc être analysé pour étudier l'intérêt conjoint des classifieurs fournis en entrée.
Un exemple simple d'opérateur d'agrégation est la fonction $\mathrm{Min}(p_1,p_2)$: soit p_1 et p_2 les probabilités d'appartenance d'une séquence à la famille de protéines telles qu'elles ont été estimées par les classifieurs c_1 et c_2, l'appartenance de la séquence à la famille sera définie comme la valeur minimale de ces deux probabilités. Ce type d'opérateur est une technique d'agrégation « prudente » : elle considère systématiquement l'avis du classifieur le plus pessimiste.

Un essaim grammatical [O'Neill 2006] a été conçu pour générer automatiquement des opérateurs d'agrégation. L'évolution grammaticale [O'Neill 2003] (*Grammatical Evolution* ou GE par la suite) est une méthode évolutionnaire dont le but est de déterminer un programme possédant un certain degré d'adaptation. A l'instar de la Programmation Génétique, un GE recherche automatiquement un programme π qui réalise "au mieux" un calcul. La solution d'un GE est donc généralement une *fonction*, et non un ensemble de données optimisant une fonction objectif. Un GE tout comme un algorithme génétique est un problème d'optimisation : on cherche à optimiser l'adaptation d'une fonction à un problème donné. Notons que le terme programme est dans ce contexte synonyme de fonction. Dans notre cas précis, l'opérateur d'agrégation est la fonction (ou programme) recherchée.

La fonction décrite par ce programme est générée à partir d'un ensemble d'opérateurs arithmétiques et logiques élémentaires. Considérons l'exemple de la recherche d'un programme fournissant une réponse *z* à une entrée (*x,y)*, les variables *x*, *y* et *z* étant entières. Connaissant un ensemble de valeurs d'apprentissage, le GE doit établir la fonction qui se rapproche au mieux de ces valeurs. Considérons par exemple les triplets (*x,y,z*) suivants : (1,0,0), (1,3,3), (2,1,2), (2,3,6). Une solution optimale pour ces données est la fonction $z=xy$.

La technique des GEs a été utilisée dans des applications financières telles que la prédiction de faillites de société et en bio-informatique [O'Neill 2005]. GE se distingue des techniques usuelles

de Programmation Génétique par l'emploi explicite d'une grammaire définissant la structure des programmes générés. GE a donc l'avantage de garantir un résultat correct d'un point de vue syntaxique. Cette particularité des GEs permet aussi de contraindre l'espace de recherche en jouant sur la définition de la grammaire, en autorisant ou non telle règle de production syntaxique. Par ce fait, on peut contraindre un GE de produire des solutions compatibles avec des connaissances a priori.

La puissance de GE repose sur la distinction opérée entre le "génotype" et le "phénotype", offrant ainsi une remarquable flexibilité. Pour illustrer cette propriété, reprenons l'exemple précédent basé sur les variables (x,y,z). Dans notre contexte, le phénotype est une fonction candidate, comme par exemple la fonction $z=x+y$. Ce phénotype a été établi à partir d'un génotype spécifique à cette fonction. Le génotype est déterminé par un "code génétique", ce dernier prenant généralement pour un GE la forme d'une vecteur numérique. Supposons que ce code soit réduit à une seule valeur entière comprise entre 1 et 4. La transformation du code génétique en un programme (phénotype) est déterminé par un ensemble de règles grammaticales. Dans notre exemple simplifié (un seul gène), une seule règle est établie. Celle-ci permet de sélectionner une des quatre opérations arithmétiques (les opérandes étant x et y) : la valeur 1 désigne l'addition, 2, la soustraction, 3, la multiplication et 4, la division. Dans ce modèle, il n'existe que 4 types distincts d'opérateurs, les fonctions les plus aptes correspondent donc au génotype 3. Cet exemple est évidemment très simpliste. Le chapitre 3.1 développe plus en détails la méthode utilisée, mais d'ores et déjà il est utile de noter la distinction existante entre un génotype et son phénotype.

Cette distinction implique que l'utilisateur peut modifier aisément la structure des programmes en changeant la spécification du langage, sans avoir à modifier l'algorithme de recherche. Cette propriété est un bénéfice certain par rapport à la programmation génétique. Dans cette dernière, le programme, représenté par une expression arborescente, est directement manipulé par des opérateurs dédies. En GE, les opérateurs génétiques modifient des vecteurs numériques. Ces derniers sont ensuite traduits en un programme en utilisant une grammaire.

Le fait d'utiliser une grammaire possède un grand avantage : on dispose ainsi d'une représentation formelle d'un langage sous la forme d'un corpus de règles explicites qu'on peut réduire ou enrichir sans remettre en cause l'algorithme évolutionnaire.

Ce découplage implique que cette technique peut être associée à n'importe quelle méthode évolutionnaire, l'approche la plus classique étant l'emploi d'algorithmes génétiques. L'essaim grammatical (*Grammatical Swarm* ou GS par la suite) est une variante récente de GE, axée sur l'optimisation par essaim de particules (Particle Swarm Optimization ou PSO par la suite ; [Kennedy 1995]). Cette méthode est une optimisation stochastique inspirée par le comportement social de nuées d'oiseaux ou de bancs de poissons.

Par comparaison aux algorithmes génétiques qui utilisent des opérateurs coûteux en temps d'exécution (les opérateurs de mutation et de croisement), PSO apparaît comme une méthode originale et efficace. Il a été montré expérimentalement que cet algorithme nécessite des populations de taille plus petite que celles des algorithmes génétiques, tout en maintenant une taille fixe des individus.

3.1. Principe général de l’essaim grammatical

La Fig. 10 présente le principe général gouvernant un GS. La méthode repose sur une correspondance entre génotype et phénotype. Le génotype est défini par une liste ordonnée d'entiers codant le programme. Ces valeurs numériques déterminent les règles de production à employer (règles exprimées dans la forme Backus-Naur). Cette transformation induit une séquence de dérivation qui à son tour engendre le phénotype, à savoir le programme représenté par une structure arborescente à l'instar de la technique de programmation génétique.

La forme Backus-Naur est une notation spécifiant un langage en termes de règles de production. La Fig. 11 indique la grammaire que nous avons conçue pour notre opérateur d'agrégation. Le symbole <S> représente le symbole d'entrée de la grammaire. En plus des habituels opérateurs arithmétiques, nous avons introduit deux opérateurs fréquemment employés pour l'agrégation de classifieurs : les opérateurs Min(a,b) and Max(a,b). Ces opérateurs renvoient la valeur minimale (resp. la valeur maximale) de a et de b.

Nous avons également défini un opérateur analogue à celui existant dans certains langages de programmation comme le langage C. Cet opérateur intitulé CND permet de spécifier des instructions conditionnelles. L'opérateur (CND ? i1 : i2) évalue la condition CND. Si celle-ci est vraie, l'instruction i1 est prise en compte, autrement l'instruction i2 est utilisée.

Le seul opérateur logique du langage est l'opérateur Less (opérateur <), mais des opérateurs complexes peuvent être générés par imbrication d'opérateurs conditionnels. On note que GS peut utiliser l'opérateur minimum ou le définir à travers l'expression ((Less x y) ? x ; y). Cette redondance logique existe aussi dans des langages de programmation comme par exemple le langage Java (choix entre la fonction Math.min(a,b) et une instruction conditionnelle).

Les symboles terminaux sont associés aux entrées, qui sont les résultats des classifieurs (estimation de probabilité d'appartenance à la classe positive). Un symbole terminal additionnel a été conçu pour permettre la déclaration de paramètres supplémentaires, tels que des seuils générés aléatoirement. Le principe de création de ces valeurs sera explicité dans la section suivante.

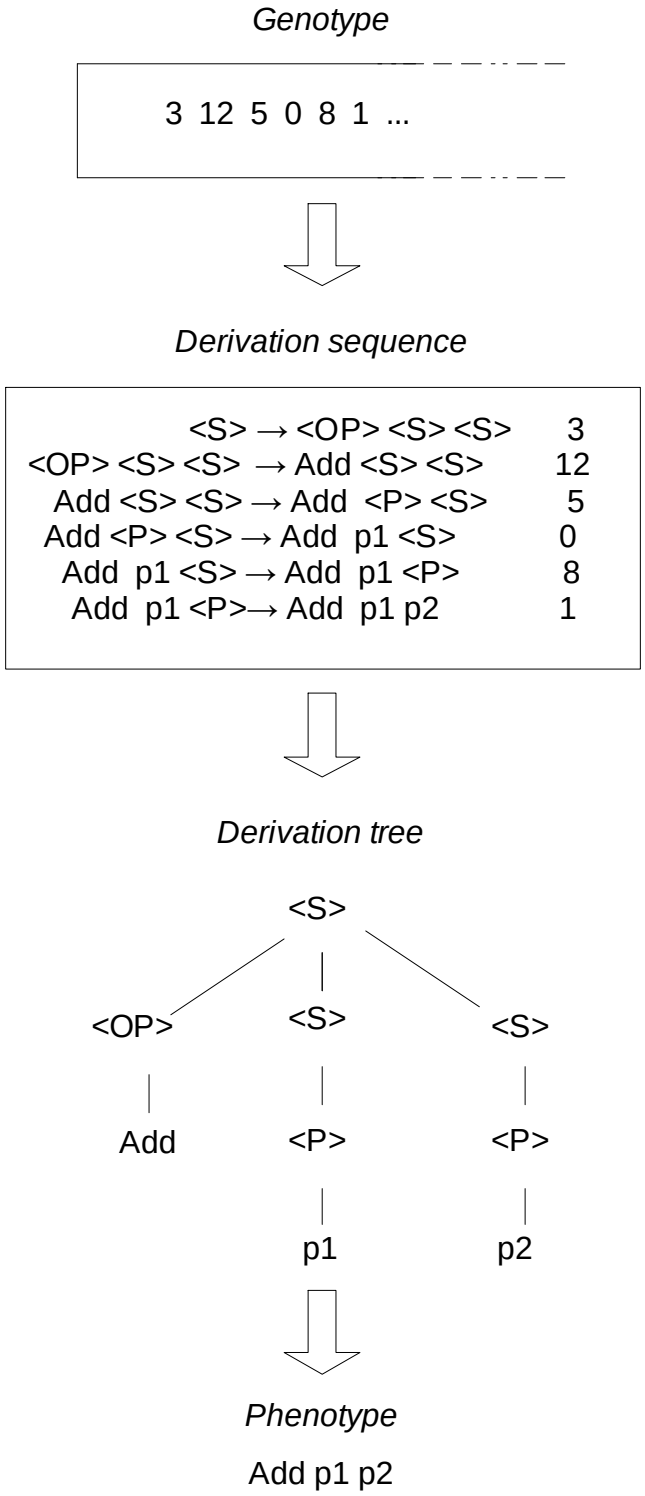


Fig. 10. *Exemple d'association génotype/phénotype.*
Le génotype est codé sous forme d'un vecteur numérique. La séquence de dérivation détaille la transformation en règles de production déterminant l'arbre de dérivation. Ce dernier peut être simplifié et représenté par son expression syntaxique (phénotype résultat).

<S>	::=	<Op> <S> <S>	(1.0)
	|	(<CND> ? <S> : <S>)	(1.1)
	|	<P>	(1.2)
<Op>	::=	Add	(2.0)
	|	Sub	(2.1)
	|	Mul	(2.2)
	|	Div	(2.3)
	|	Min	(2.4)
	|	Max	(2.5)
<CND>	::=	Less <P> <P>	(3.0)
	|	Less <P> <T>	(3.1)
	|	Less <T> <P>	(3.2)
<T>	::=	τ	(4.0)
<P>	::=	p1	(5.0)
	|	p2	(5.1)
	|	…	(5.i)
	|	pn	(5.n-1)

Fig. 11. *Grammaire générative utilisée pour la construction d'opérateurs d'agrégation. L'expression (p.q) à droite des règles de production désigne le numéro de la règle (p) ainsi que le numéro de la branche (q) de la règle.*

3.2. Génération des opérateurs d'agrégation

Le processus de transformation utilise le génotype pour construire une expression en suivant les commandes apportées par les règles de production. L'expression est créée initialement avec un seul symbole germinal, le symbole <S>. Ce dernier peut être remplacé par une des branches de droite de la règle correspondante, à savoir les branches (1.0), (1.1) and (1.2) définies dans la Fig. 11. La sélection d'une de ces trois possibilités dépend du génome de l'individu concerné. Ce génome est codé par une liste ordonnée d'entiers positifs (c_1, c_2, . . . , c_n). Soit r le nombre d'alternatives d'une règle et a le branchement retenu. Dans la version de base de GE, la valeur de a est donnée par la formule suivante :

$$a = c_i \,\%\, r \qquad (10)$$

où c_i est le $i^{ème}$ codon du génome et % l'opérateur modulo.

Sur l'exemple de la Fig. 10, la branche (1.0) est sélectionnée étant données les valeurs c_1 =3 et r =3. Dans l'expression courante, on remplace le symbole de gauche de la règle par les symboles de droite. Ainsi le symbole <S> sera remplacé par <OP> <S> <S>. La même procédure est ensuite appliquée, en considérant le premier symbole de la nouvelle expression : <Op>. La règle correspondante (2) possède 6 branches : (2.0) à (2.5). Le codon courant étant c_2=12, le symbole <Op> est remplacé par le symbole <Add>. La Fig. 10 indique l'arbre de dérivation obtenu à partir

de la lecture du génome.
Lors d'une telle transformation, deux situations peuvent se présenter : soit tous les symboles non terminaux ont été remplacés dans l'expression et le programme est alors valide, soit le dernier codon du génome a été utilisé sans que la procédure ait pu être achevée. Dans ce dernier cas, la lecture recommence depuis le début du génome.

La règle 4 (Fig. 11) est un artifice pour créer des valeurs aléatoires de seuil. Comme la règle ne possède pas d'alternative, le codon courant est disponible. Il est utilisé pour créer un seuil comme dans l'exemple suivant : ((Less 0.9 p1) ? p1 : p2). Cette règle édicte que si p1 > 0.9 alors la valeur retenue est p1 sinon p2. La valeur 0.9 est donnée par la valeur du codon (pour ramener cette valeur entre 0 et 1 on applique une normalisation par *cmax* (*cmax* étant la borne supérieure du codage entier des codons).

Une autre version de GS a également été implémentée, dans laquelle le choix de la branche n'est pas déterminée par l'opérateur modulo. Une critique en effet de cette stratégie est qu'elle accorde la même importance (et donc la même probabilité d'apparition) à toutes les branches. La version proposée établit des poids pour chaque branche, comme le montre l'exemple de la Fig. 12 : la règle X ::= Y a une probabilité de 0.25 d'être retenue, le choix étant établi par la valeur de ci. Ce principe est analogue à celui de la roue de la fortune utilisée dans certains algorithmes génétiques.

X	::=	Y	$c_i \in [0,63]$
	|	Z	$c_i \in [64,255]$

Fig. 12. *Règle de production probabiliste.*

3.3. Optimisation par essaim de particules

PSO est un algorithme inspiré des modèles sociaux de comportement, comme le vol d'essaims d'oiseaux par exemple. Cet algorithme simule les réactions des animaux sociaux dans leur groupe, chaque individu guettant ce que font les autres, tiraillés entre leurs espérances de gain et celles de ses congénères. PSO est basé sur une population d'acteurs dénommés particules. Un essaim de particules explore un espace de recherche à *n* dimensions. La position d'une particule détermine une solution potentielle. Chaque particule connaît la meilleure position (*pbest*) qu'il a rencontrée jusqu'ici ainsi que la meilleure position (*gbest*) obtenue par ses voisins (dans notre implémentation le voisinage englobe l'ensemble des particules). Les particules ajustent leur trajectoire en fonction de leur propre expérience et celles de leurs voisins. Le déplacement d'une particule est défini par sa position courante ainsi que par un vecteur de vitesse mis à jour à chaque itération de l'algorithme. L'algorithme débute avec une population composée de particules ayant des positions et des vitesses aléatoires. Une valeur d'adaptation est alors estimée pour chaque particule : les valeurs de *pbest* et *gbest* sont réactualisées si nécessaire. Le vecteur de vitesse v_i à l'itération *t+1* est déterminé par la formule suivante :

$$v_i(t+1) = w.v_i(t) + \gamma_1.R_1.(pbest - x_i) + \gamma_2.R_2.(gbest - x_i) \quad (11)$$

où

$$w = \text{wmax} - ((\text{wmax} - \text{wmin}) / \text{itermax}).\text{iter} \quad (12)$$

Les poids γ_1 et γ_2 sont associés respectivement aux termes individuel (*pbest*) et social (*gbest*). R_1 et

R_2 sont des valeurs aléatoires uniformes entre 0 et 1. x_i désigne la i^{ème} composante du vecteur position de la particule. Le paramètre *w* est un coefficient décroissant temporellement de wmax à wmin pour accorder une liberté de recherche plus grande au début (trajectoire aléatoire pour explorer l'espace sans a priori).

Une fois le vecteur de vitesse réactualisé, on détermine la nouvelle position de chaque particule par la formule suivante :

$$x_i(t+1) = x_i(t) + v_i(t+1) \qquad (13)$$

L'espace de recherche étant borné, une composante hors-limite est prohibée ; la valeur est ramenée à la valeur minimale ou maximale autorisée, selon le cas considéré.

3.4. Détermination de la fonction objectif

Comme énoncé à la fin du chapitre précédent, l'aire sous la courbe ROC ou AUC est une mesure pertinente pour quantifier la qualité d'un classement. La fonction objectif du GS (le terme adaptation est également employé) peut donc reprendre le calcul de l'AUC. Nous avons cependant introduit un facteur d'ajustement ϕ pour contrôler la structure du programme généré par l'algorithme GS.
La fonction objectif associée à un programme π est définie par la formule suivante :

$$\zeta(\pi)=\phi(\pi).AUC(\pi) \qquad (14)$$

Le but de ce facteur d'ajustement est de contraindre l'essaim de particules à explorer certaines régions privilégiées de l'espace de recherche, afin de maîtriser la complexité du programme. Cette dernière peut être modélisée par le nombre *c* de classifieurs pris en compte ainsi que par le nombre *o* d'opérations utilisées : $\phi(\pi) = \phi(c, o)$. Le facteur d'ajustement d'un programme π est donc une fonction prenant en compte la structure du programme, et plus précisément sa complexité, exprimée par le couple de paramètres (c, o).
Par exemple, le programme π = (Add (Min p1 p2) p1) présente une structure (c=2, o=2). Tous les opérateurs sont comptés pour une unité d'opérations, seul CND, opérateur complexe impliquant 4 opérandes, équivaut à deux unités. Ainsi le programme π = ((Less p1 p2) ? Min(p1,p3) : p2) possède une structure (c=3, o=3).

Nous proposons le facteur d'ajustement suivant :

$$\phi(c, o) = 1/(1 + \exp(dist(c, o) - \delta)) \qquad (15)$$

Ce facteur d'ajustement est une fonction à base radiale illustrée par la Fig. 14. La valeur maximale de la fonction correspond à un point central C de coordonnées (x_c,y_c), indiqué sur la Fig. 14, et correspondant à une complexité moyenne souhaitée par l'utilisateur. Lorsque la complexité d'un programme (exprimée par le couple de paramètres (c,o)) s'écarte de cette zone centrale, la valeur du facteur d'ajustement décroît, introduisant ainsi une pénalité au programme π correspondant. Le contrôle de cette décroissance est établie au moyen d'une fonction dist explicitée sur la Fig. 13, ainsi que par un paramètre de translation δ (nous avons utilisé une fonction sigmoïde plutôt qu'une gaussienne pour avoir un décroissement modéré aux alentours du point central).

Ce facteur d'ajustement résulte d'un principe de parcimonie : on souhaite garantir une complexité raisonnable de l'opérateur. Le nombre d'opérateurs doit couvrir une étendue restreinte pour éviter les problèmes de sur-apprentissage. Par ailleurs, le nombre de classifieurs n'est pas indifférent : la confiance accordée à une fusion de classifieurs dépend du nombre d'experts donnant leur avis. Un nombre trop faible n'est pas une stratégie intéressante (on sait qu'aucun classifieur ne surpasse réellement les autres). Un nombre trop grand impliquerait par ailleurs une plus grande complexité calculatoire pour le système de classification final : il est avantageux d'exclure certains classifieurs dont l'apport est trop faible sur la performance globale. On sait que certains classifieurs sont assez proches par nature (comme Spectrum Kernel et Mismatch par exemple) et peuvent s'avérer redondants.

dist(*c*,*o*) : renvoie une valeur réelle comprise entre 0 et 1
paramètres :
x_c, y_c : coordonnées d'un point central C
σ_x, σ_y : poids
début
$x = c - x_c$;
$y = o - y_c$;

$$x' = \frac{\sqrt{2}}{2}(x - y) ;$$

$$y' = \frac{\sqrt{2}}{2}(x + y) ;$$

$$dist = \sqrt{\frac{x'^2}{2\sigma_x^2} + \frac{y'^2}{2\sigma_y^2}} ;$$

fin.

Fig. 13. *Calcul de la fonction dist.*
La rotation de 45 degrés opérée permet d'orienter la fonction sur la bissectrice. Cette rotation repose sur le postulat que les deux paramètres c *et* o *augmentent proportionnellement.*

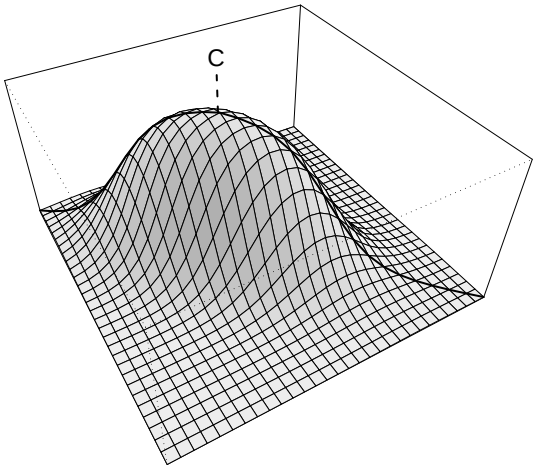


Fig. 14. *Vue en perspective du facteur d'ajustement.*
Les paramètres que nous avons définis dans notre implémentation ceux de la fonction dist (coordonnées du point central C (x_c = 4, y_c = 4) et poids σ_x=2, σ_y =2 ; voir Fig. 13), ainsi que le paramètre de translation δ=3.

3.5. Les "classifieurs experts" pour l'identification d'homologues

Le système de classification repose sur l'avis donné par un nombre *nc* de "classifieurs experts". Ces derniers sont de deux types. Le premier type est donné par des classifieurs SVM connus pour leurs performances en classification de protéines. Nous avons repris les classifieurs connus pour leurs bonnes performances en classification de protéines : SpectrumKernel, Mismatch, Pairwise, PairwiseBlast et LAKernel. Les classifieurs SVM possèdent une capacité remarquable à traiter des données exprimées dans des espaces de grande dimension. Ce type de classifieur peut être qualifié de généraliste dans le sens où il prend en compte toute l'information de la séquence. A ce groupe d'experts s'ajoute une deuxième catégorie plus spécialisée, chacun de ses membres apportant une information biologique supplémentaire. Contrairement à un classifieur du premier groupe, ces classifieurs experts ne peuvent à eux seuls identifier la classe d'une protéine, mais ils apportent un indice complémentaire pouvant améliorer les performances du système global de classification. Nous avons défini les classifieurs experts suivants :

- la taille de la séquence. Bien que ce critère semble simpliste, la taille peut partiellement séparer les cytokines des autres protéines. On sait en effet que les cytokines ont une variabilité de taille faible par rapport à l'ensemble du protéome.
- le point isoélectrique. Ce paramètre est défini par le PH pour lequel une molécule est dans l'état zwitterionique (un état caractérisé par l'absence de charge électrique). Les cytokines possèdent des comportements biologiques similaires ; ils doivent partager certaines propriétés physico-chimiques.
- le critère SOV (*segment overlap*). La structure secondaire des cytokines est très bien conservée. Nous avons utilisé le logiciel PSIpred [Jones 1999] pour prédire la structure secondaire des séquences protéiques et le critère SOV [Zemla 1999] pour quantifier la similarité de structures secondaires entre deux séquences.

Tous ces classifieurs experts ont été modélisés sous la forme de classifieurs Bayésiens naïfs.

3.6. Tests de performance

La famille des cytokines comprend 45 séquences et les contre-exemples ont été extraits de la base SCOP. Comme le montre la Fig. 15, trois jeux de données ont été sélectionnés. Le premier jeu S_1 est utilisé pour entraîner les classifieurs, un deuxième jeu d'apprentissage S_2 est appliqué pour entraîner l'opérateur d'agrégation et le troisième jeu S_3 permet d'évaluer les performances du classement. Le jeu de cytokines a donc été divisé en 3 sous-ensembles de taille 15 (exemples positifs de S_1, S_2 et S_3). L'ensemble S_1 comporte 100 contre-exemples : les membres restent inchangés pour ne pas influencer le résultat final (la capacité d'apprentissage des classifieurs demeure constante). Les deux autres jeux S_2 et S_3 sont également composés de 15 séquences positives et 100 contre-exemples. Ces derniers jeux font l'objet de la validation croisée.

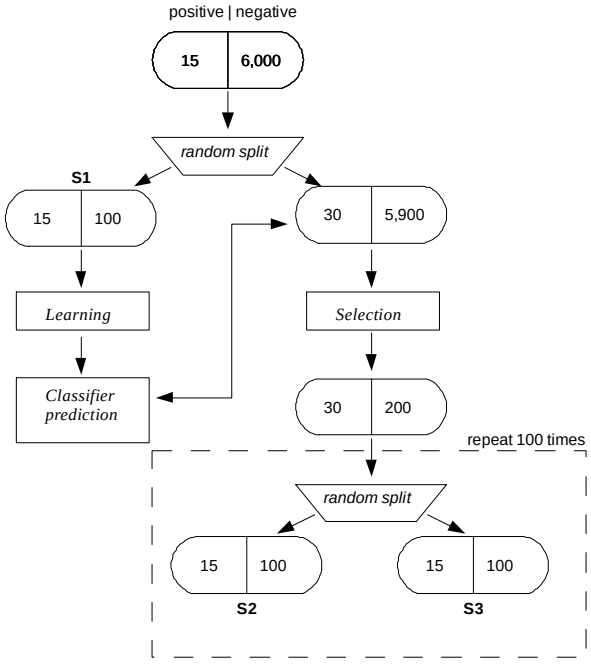


Fig. 15. *Principe de construction des jeux de données.*
La méthode Random split définit l'ensemble d'apprentissage S1 utilisé par les classifieurs. Une phase de prédiction fournit ensuite les estimations de probabilité servant d'entrées pour le GS. La méthode Selection extrait les 200 séquences les plus difficiles à classer par les classifieurs (40 premiers rangs d'exemples négatifs pour chaque classifieur). Enfin, la validation croisée est utilisée pour définir des jeux d'apprentissage et de tests.

La Fig. 16 représente les scores AUC obtenus par validation croisée. Pour des raisons de comparaison de performances, nous avons implémenté une règle triviale d'agrégation appelée BEST qui ne retient qu'un seul avis : celui du meilleur classifieur sur le jeu d'apprentissage S_2. Ce meilleur classifieur est ensuite utilisé pour classer le jeu de test S_3.

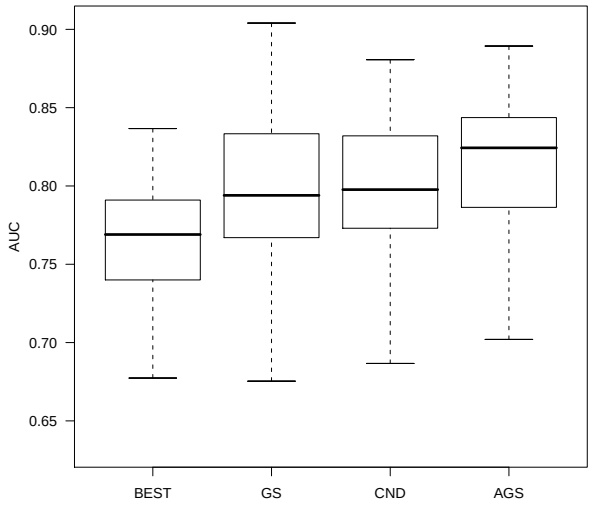


Fig. 16. *Comparaison de performances de classification.*
Cette figure représente les scores AUC pour la méthode BEST ainsi que plusieurs implémentations de GS. Contrairement à AGS et GS qui acceptent tout type d'opérateurs, CND est uniquement basé sur des opérateurs conditionnels. L'adaptation de GS est le score AUC tandis que celui de la méthode AGS utilise le facteur d'ajustement pour adapter le score en fonction de la structure du programme généré.

Les résultats montrent que la méthode qui retient le meilleur classifieur n'est pas optimale, ce qui indique une certaine complémentarité des experts. On remarque la bonne performance de CND, pourtant basée sur un seul opérateur conditionnel. L'analyse des opérateurs d'agrégation générés par la méthode GS montre en effet que les programmes obtenant le score le plus élevé d'adaptation sont ceux possédant une structure simple, impliquant peu d'opérateurs et peu de classifieurs. Ce phénomène justifie le principe de parcimonie guidant la méthode AGS. Cette dernière, par le biais de son facteur d'ajustement, surpasse en effet les performances de la version GS de base.

La Table 8 présente quelques exemples de programmes générés par l'essaim grammatical. La plupart des résultats semble indiquer que seule la moitié des classifieurs SVMs sont nécessaires pour obtenir de bonnes performances, à condition d'ajouter des experts. Ce résultat peut sembler surprenant, mais il peut s'expliquer par une certaine redondance des classifieurs entre eux (leurs résultats sont similaires et leur utilisation conjointe n'augmente pas les performances) et par la complémentarité des experts.

Opérateur d'agrégation	*Score AUC*
(Mul (Min p6 p8) (Min p1 p4))	0.86
((Less 0.70 p4) ? p5: (Min p8 p6))	0.88
(Add (Add p6 p7) (Less 0.61 p1) ? p4: p1))	0.90
((Less p4 p2) ? p1 : (Add p6 ((Less 0.84 p4) ? p1: p7)))	0.91

Table 8. *Exemples d'opérateurs d'agrégation.*
Les opérandes pi sont les estimations de probabilité données par les classifieurs SVM (i de 1à 5) et les experts Bayésiens (i de 6 à 8), les indices correspondant à l'ordre de présentation de ces méthodes.

3.7. Conclusion

Cette étude a montré l'intérêt de combiner plusieurs classifieurs pour accroître les performances d'identification d'homologues. Si on considère en effet le classifieur ayant obtenu la meilleure performance en apprentissage, sa performance en test est largement inférieure à celle obtenue par notre méthode d'agrégation.
Les tests ont été menés sur la base du critère AUC et non seulement sur le taux d'erreurs. Obtenir un meilleur classement est en effet une nécessité cruciale, étant donné le grand nombre de séquences candidates et le caractère fastidieux d'une validation par un biologiste. Ces résultats révèlent la pertinence des experts automatiques proposés. Ils apportent une puissance de discrimination supérieure si on les applique conjointement à des classifieurs SVM. L'approche par essaim grammatical possède l'avantage de fournir un algorithme de classement compréhensible par l'utilisateur et de mettre en évidence l'intérêt relatif de chacun des classifieurs.

Mon apport dans cette thématique est fortement lié à la méthode utilisée. La technique des essaims grammaticaux est en effet relativement récente et peu de travaux ont été publiés dans des applications autres que des exemples jouets. De la même manière, il n'y a pas eu à ma connaissance d'évaluation des avantages des essaims grammaticaux par rapport aux évolutions grammaticales. J'ai montré dans mon application que la méthode des essaims grammaticaux donne des résultats légèrement meilleurs que ceux obtenus sur une version de base d'algorithme génétique (sélection par tournoi, mutation et croisement sur une population unique). Dans notre contexte, les essaims de particule ont été plus performants que cet algorithme génétique. Les avantages par rapport aux algorithmes génétiques sont l'utilisation d'une population réduite, une convergence rapide ainsi que l'absence d'opérateurs complexes.

Partie B. Analyse de données de puces à ADN

La deuxième partie de ce mémoire porte sur l'analyse du transcriptome. Le quatrième chapitre expose les techniques des puces à ADN ainsi que les mesures associées. Le chapitre suivant est consacré à l'analyse implicative des données d'expression. Cette analyse est une application directe de l'Analyse Statistique Implicative à l'étude des données du transcriptome. Elle comprend notamment des méthodes originales pour l'extraction de gènes discriminants ainsi que pour la classification de tumeurs. Enfin, le sixième chapitre présente une application de méta-analyse de puces à ADN.

Chapitre 4. Le transcriptome

Le transcriptome est l'ensemble des *ARN messagers* ou *transcrits* d'un organisme vivant et son étude est une mine précieuse d'informations sur l'expression du génome. L'analyse du transcriptome vise donc à connaître l'activité cellulaire en s'intéressant aux gènes exprimés dans un contexte expérimental donné. Il est important de noter que le transcriptome décrit l'activité d'un type cellulaire donné. Chez l'homme, sur environ 200 000 ARNm transcrits, seuls 10 000 à 20 000 sont exprimés dans une cellule spécialisée [Le Meur 2004]. De plus, parmi ces transcrits, 4 000 à 6 000 semblent spécifiques de ce type cellulaire.

L'analyse du transcriptome permet d'appréhender les mécanismes de régulation des gènes et d'établir des réseaux d'interaction de gènes. On comprend dès lors qu'à travers le transcrit, c'est le gène qui est l'objet d'intérêt du biologiste : par la suite on utilisera par abus de langage le terme de gène en lieu et place de celui de transcrit.

Le bon fonctionnement d'une cellule et donc d'un organisme vivant dépend de la coopération de plusieurs milliers de gènes par le biais de leurs produits d'expression (ARN et protéines). Toute anomalie dans la séquence génétique, mais aussi dans la régulation des gènes peut être source de nombreuses pathologies. Les premières techniques d'analyse portaient sur l'étude d'un seul gène (ou d'un nombre restreint de gènes) pour une expérimentation donnée. Des approches telles que le *Southern blot* et le *Northern blot*, permettent d'identifier et de localiser une séquence particulière (sonde d'ARNm ou ADNc) dans un génome entier (cible). Ces techniques procèdent par une fragmentation des acides nucléiques, suivie par une électrophorèse (séparation des fragments selon leur charge électrique).
La technologie des puces à ADN a considérablement bouleversé le travail des biologistes en apportant la possibilité d'analyser l'expression simultanée d'un grand nombre de gènes, voire de l'ensemble des gènes contenus dans le génome. Cette approche dite : « à haut débit » a ouvert la voie à une nouvelle discipline, la génomique fonctionnelle, le chercheur pouvant désormais étudier le comportement coordonné des gènes et ainsi mieux appréhender la fonction d'un gène dans une situation expérimentale précise.

L'étude des pathologies humaines est une application majeure de cette technologie, en cancérologie notamment (on peut citer les lymphomes [Bertucci 2002], les leucémies [Bullinger 2005], le cancer du sein [Bertucci 2005], de la prostate [Lapointe 2004] et le mélanome [Pavey 2004]).
L'utilisation des puces à ADN a permis d'améliorer la classification des tumeurs ([Alizadeh 2000], [Thieblemont 2004], [Ballester 2005]), elle a aussi affiné la définition de celles-ci, en établissant des sous-types tumoraux qui étaient difficilement analysables par les voies traditionnelles utilisées jusqu'alors par les cliniciens. Elle a également mis en évidence de nouvelles cibles thérapeutiques et a permis de prédire la survie des patients.

4.1. Définition et principe des puces à ADN

Une puce est un support plan, qui peut être en verre, silicium ou plastique, sur lequel vont être fixées des molécules. Ces molécules sont le plus souvent des acides nucléiques : *ARN*, *ADN*, *ADNc*, *ARNc*. Il existe toutefois d'autres types de puces comme notamment les puces à protéines. Outre le type de support et la nature de l'ADN, les puces peuvent être classées selon le nombre de sondes qu'elles contiennent. Les sondes correspondent aux molécules fixées sur la lame, chaque sonde

étant propre à une séquence particulière. Le principe général des puces est basé sur une propriété fondamentale qui est l'hybridation des acides nucléiques. Cette dénomination recouvre la capacité des acides nucléiques simples brins à s'associer pour former une molécule double brin. Une expérimentation part d'une population d'ADN qui est le sujet de l'étude (extrait d'un tissu pour un diagnostic par exemple). Cette population constitue les séquences cibles à comparer avec les séquences des sondes de la puce. L'hybridation est la clé de la reconnaissance des gènes exprimés, les cibles allant s'apparier avec les sondes dont la séquence est compatible (par un principe de complémentarité des bases), comme le montre la Fig. 17.

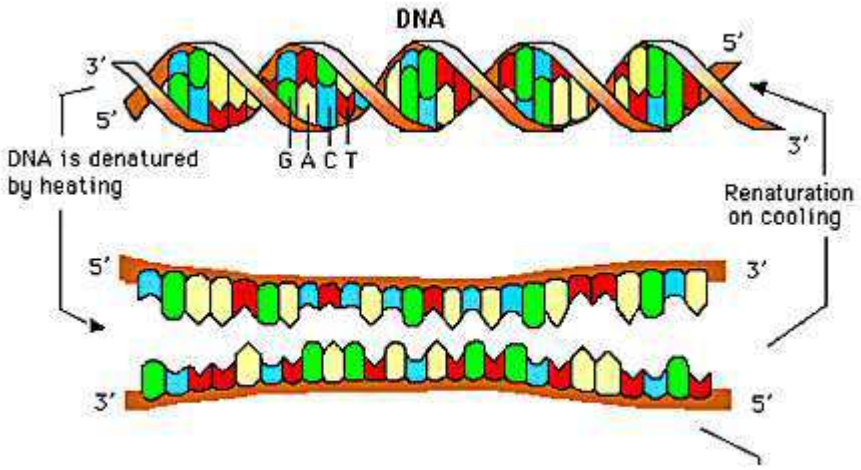


Fig. 17. *Complémentarité des brins d'ADN.*
L'ADN est structurée en une double hélice, chaque brin pouvant être décrit par une séquence composée de quatre bases (l'adénine (notée A), la thymine (notée T), la cytosine (notée C) et la guanine (notée G)). Un brin s'apparie avec son brin complémentaire grâce à des liaisons hydrogène formées entre les bases A-T d'une part, et C-G d'autre part.

Une puce servant pour le diagnostic pourra ne contenir qu'une vingtaine de transcrits très ciblés, alors qu'une puce pangénomique pourra regrouper des dizaines de milliers de transcrits (à titre d'exemple, la puce Genechip® *human genome U133 Plus 2.0 array*, commercialisée par Affymetrix, permet l'analyse simultanée de plus de 47 000 transcrits). On distingue les puces spottées et les puces à *oligonugléotides*. Pour les premières, les sondes sont d'abord synthétisées puis déposées sur une lame de verre ; le moyen le plus généralement utilisé est un robot spotteur équipé de pointes qui déposent le produit selon un arrangement régulier (d'où le nom de *microarray*). Cette technique permet de concevoir des puces « maison » : les puces peuvent être facilement adaptées en sélectionnant les sondes et leur positionnement sur la lame en vue d'une expérimentation précise. Cette méthode, à condition de disposer de l'équipement approprié, est moins coûteuse qu'une puce commerciale. Cette approche est utilisée lorsqu'on connaît les gènes d'intérêt pour une étude donnée. Le deuxième type de puces sont les puces à *oligonucléotides*, ces dernières étant synthétisés *in situ* par une technique de photolithographie. Le haut degré de miniaturisation offre la possibilité d'étudier un génome entier à partir d'une seule puce.

Les puces se classent également en deux types selon qu'elles soient à simple canal ou double canaux. Dans le premier cas, on mesure directement le niveau d'expression d'un gène tandis que dans le deuxième on analyse le niveau d'expression dans deux contextes comme par exemple un tissu malade par rapport à un tissu sain servant de référence. La technologie qui permet ce type d'étude est basée sur deux marquages fluorochromes (un canal rouge et un canal vert) et une analyse d'image par un scanner. La mesure est généralement le rapport logarithmique des intensités dans les deux canaux, définissant le niveau d'expression relatif d'un gène dans une situation par rapport à une autre (comme par exemple la surexpression d'un gène dans une tumeur par rapport à

son expression dans un tissu sain).

La Fig. 18 rappelle la technique mise en œuvre pour des puces spottées et la Fig. 19 donne un exemple de résultat d'image obtenue à partir d'une puce.

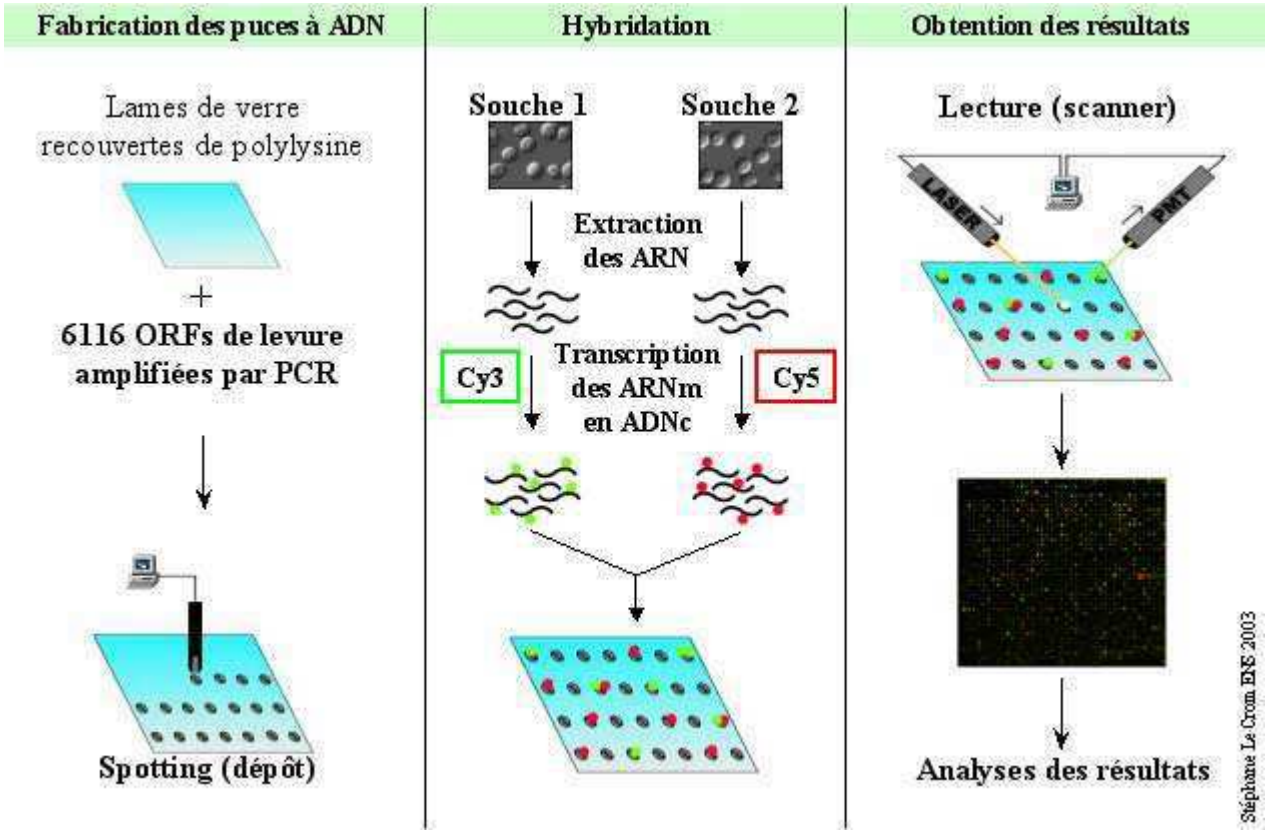


Fig. 18. *Principe des puces à ADN*
La mesure d'expression des gènes procède en trois étapes. En premier lieu, la fabrication de la puce consiste à déposer des sondes sur la lame (figure de gauche). La deuxième étape est la technique d'hybridation, qui va mettre les cibles au contact des sondes. Dans l'exemple (figure du milieu), la technique utilise deux canaux (marquage par les fluorochromes Cy3 et Cy5). La dernière étape est l'acquisition par scanner d'une image de la puce. Cette image est analysée pour détecter les spots et quantifier les niveaux d'expressions dans les deux canaux.

Il est important de noter que les puces à ADN mesurent des niveaux d'hybridation (et non des quantités de séquences d'ARNm) moyennés sur des populations de cellules (et non une cellule unique). Le « dogme central » des puces à ADN est que le niveau d'expression des ARNm est corrélé positivement avec celui de la protéine associée. Il existe cependant bien d'autres facteurs intervenant dans l'expression d'une protéine.

Les puces *GeneChip* de la société Affymetrix sont les plus largement répandues en analyse du transcriptome. Le principe de mesures de ces puces mérite une présentation détaillée afin de saisir toutes les difficultés de ce type d'analyse. Les puces Affymetrix sont des puces mono-couleur pour lesquels chaque gène est représenté par un ensemble de sondes (*probes*) appelées "*probe sets*" (voir exemple de la Fig. 20). Les sondes possèdent 25 bases nucléiques. Un *probe set* consiste en plusieurs paires de sondes (de 10 à 16 selon la technologie). Chaque paire de probes contient une sonde dénommée *perfect match* (PM) et une sonde dénommée *mismatch* (MM). Les sondes PM et MM sont des séquences identiques à l'exception d'une erreur (*mismatch*) sur MM en milieu de séquence (base complémentaire à celle de PM). Les sondes MM sont donc conçues pour mesurer le niveau d'intensité du fond pour la sonde PM. Plus précisément, les sondes MM mesurent le

phénomène de cross-hybridation, à savoir une hybridation par une séquence partiellement complémentaire. Ce phénomène dépend de la composition de la séquence. Il arrive cependant que la sonde MM présente une intensité supérieure à celle de la sonde PM correspondante. Deux raisons peuvent intervenir : soit la séquence MM possède une forte homologie avec une séquence inconnue, soit l'échantillon présente une variation de séquence causant une meilleure hybridation. La valeur du signal est définie par une méthode complexe assurant la robustesse statistique de la mesure [Hubbell 2002] et dans laquelle intervient la différence des valeurs logarithmiques d'intensité PM et MM.

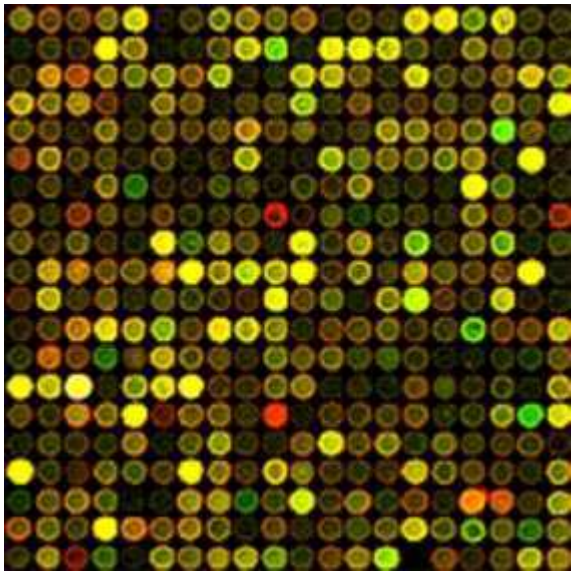


Fig. 19. *Exemple de puce à ADN*
Image en fausses couleurs obtenue à partir de l'acquisition du niveau de fluorescence dans les deux canaux. On aperçoit certains spots où une condition prédomine (en vert ou en rouge selon la condition). Les spots jaunes indiquent un équilibre entre les deux niveaux d'expression dans les deux conditions expérimentales.

gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaa**gtcaaccatccaaaactctgtttgt**caaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaa**aactctgtttgtcaaaggtctgtct**gaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtca**aaggtctgtctgaggataccactga**agagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggt**ctgtctgaggataccactgaagaga**ccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggat**accactgaagagaccttaaaagaat**catttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaa**agaatcatttgagggctctgttcgt**gcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcat**ttgagggctctgttcgtgcaagaat**agtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgag**ggctctgttcgtgcaagaatagtca**ctgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctct**gttcgtgcaagaatagtcactgatc**gggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaata**gtcactgatcgggaaactggttctt**

Fig. 20. *Sondes Affymetrix*
Sondes pour une séquence cible de 129 bases. Les 11 sondes sont réparties le long de la séquence et sont censées être spécifique de celle-ci, à l'exclusion de toute autre séquence. Le code Affymetrix pour cet ensemble de sondes est : 1415771_at. Ce jeu de sondes correspond au gène nucléoline chez la souris.

4.2. Traitement des données de puces à ADN

Les données subissent généralement un ensemble de prétraitements au demeurant assez complexe, comme il vient d'être fait mention sur l'exemple des puces Affymetrix. Les méthodes varient selon les auteurs pour éliminer le bruit de fond, les gènes invariants ou normaliser le signal. Comme mes travaux supposent que les données ont été prétraitées de façon correcte, il n'est pas utile de s'étendre sur le contenu précis de ces étapes. La procédure suivante montre un enchaînement type, sachant que d'autres algorithmes peuvent intervenir et que certaines étapes sont optionnelles :

1. recadrage de la plage de valeurs. Le signal d'expression est borné à des valeurs min. et max. prédéfinies.
2. suppression des gènes contenant un pourcentage non négligeable de valeurs manquantes.
3. suppression des gènes dont le niveau d'expression est stable sur l'ensemble des mesures.
4. remplacement des valeurs manquantes. On recherche les *k* gènes dont le profil d'expression est le plus proche du gène considéré (algorithme KNN (*k* plus proches voisins)). On remplace la valeur manquante par la valeur médiane des niveaux d'expression de ce groupe de gènes.
5. normalisation du signal par régression locale. La technique la plus courante est LOWESS (*locally weighted scatterplot smoothing* [Yang 2002]).
6. pour les données bi-canaux : calcul du ratio et transformation logarithmique.
7. centrage médian des niveaux d'expression de chaque gène.

4.3. Techniques usuelles d'analyse de données de puces à ADN

Une expérimentation repose sur un ensemble de gènes et une série de mesures d'expression prises sur des échantillons tissulaires. L'information intéressant le biologiste concerne le profil d'expression d'un gène, c'est-à-dire le vecteur de mesures associé à un gène. La Fig. 21 donne quelques exemples de profils sur une série temporelle. On constate clairement une action coordonnée de certains gènes. Un cas particulier de profil d'expression retient l'attention de nombreux biologistes, lorsque leur étude porte sur des échantillons appartenant à des classes connues. Le chercheur s'intéresse aux gènes dont le profil possède un niveau d'expression significativement différent dans une classe par rapport à une ou plusieurs autres classes (par classe on entend un groupe d'échantillons de même type, comme un tissu tumoral ou sain).

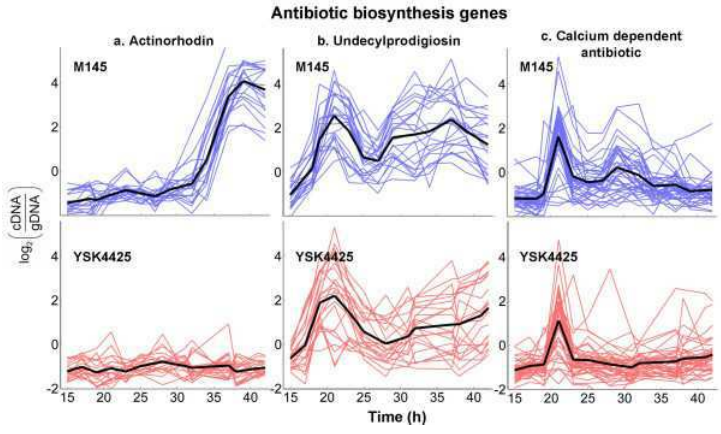


Fig. 21. *Profils d'expression*
Ces figures montrent l'évolution de l'expression des gènes lors de la biosynthèse d'antibiotiques par une bactérie dénommée Streptomyces coelicolor [Lian 2008]. L'axe des abscisses représente les temps correspondants aux mesures d'expression. L'axe des ordonnées est la valeur logarithmique relative d'expression. Le profil d'expression d'un gène désigne le vecteur de mesures d'expression relatif à un gène.

Ce type d'étude est véritablement l'application majeure des puces à ADN. Les gènes recherchés sont dits différentiels. Le test de *Student* ou t-test est souvent utilisé pour discriminer de tels gènes, avec des corrections liées au problème des comparaisons multiples [Leung 2003]. Le logiciel SAM [Tusher 2003] est très souvent utilisé pour sélectionner des gènes différentiels tout en contrôlant le taux de faux-positifs. L'analyse du transcriptome repose sur des méthodes ayant fait leurs preuves dans d'autres domaines. Ainsi, l'Analyse en Composantes Principales est employée pour visualiser et interpréter les données [Raychaudhuri 2000]. Les techniques non supervisées comme les K-means ou les Cartes auto-adaptatives [Tamayo 1999] sont également des outils retenus pour rechercher des patrons d'expression. Cependant, la méthode la plus largement répandue et servant de support à de nombreuses publications est la classification hiérarchique ascendante. Le logiciel *Cluster* [Eisen 1998] est certainement l'outil de prédilection des biologistes. Son succès tient beaucoup à la facilité d'interprétation de la représentation en fausses couleurs des données. Cet outil permet d'extraire des clusters d'expression : le terme cluster désigne un groupe de gènes ayant un profil d'expression semblable. Les clusters peuvent être extraits de manière automatique au moyen d'algorithmes dédiés. Ils sont le plus souvent sélectionnés par le biologiste lui-même. Dans la Fig. 22, les auteurs de cette étude ont isolé un cluster représentatif d'un sous-type de tumeurs faisant l'objet de leurs recherches. On distingue nettement une surexpression différenciée d'un groupe de gènes dans une classe d'échantillons. La classification hiérarchique définit un dendrogramme. Ce dernier est représenté en colonnes pour les échantillons et omis sur les gènes (en lignes) sur la figure. On remarque également sur celle-ci que la classification supervisée a « reconnu » les vraies classes d'échantillons en agglomérant prioritairement les échantillons de même catégorie. La représentation en fausses couleurs permet d'analyser aisément les niveaux relatifs d'expression. Les couleurs rouges et vertes indiquent respectivement une sur- ou sous-expression. La couleur noire signifie une expression relative quasi-nulle tandis qu'une couleur grise indique des valeurs manquantes.

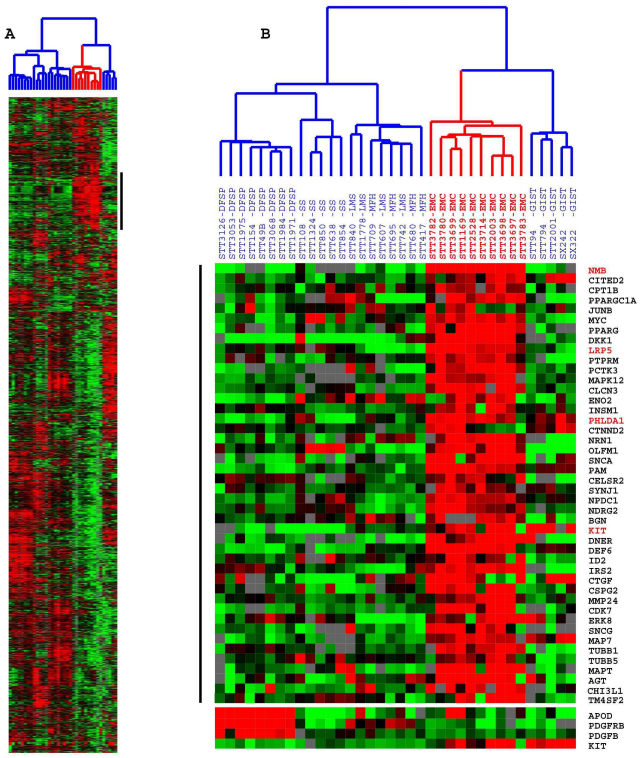


Fig. 22. *Classification hiérarchique ascendante (logiciel Cluster).*
Analyse de 26 types de sarcomes [Subramanian, 2005]. Chaque ligne représente le niveau d'expression relatif d'un gène (centrage médian). Chaque colonne représente un échantillon (36 échantillons au total). La figure A montre la totalité des mesures sur 2 918 gènes, la figure B montre le sous-ensemble de gènes différentiellement exprimés dans une classe particulière (une tumeur maligne du cartilage).

Les clusters tels que celui indiqué sur la Fig. 22 sont ensuite analysés pour établir si les gènes impliqués partagent des propriétés communes. On sait que des gènes possédant une action coordonnée participent généralement à une même fonction biologique. L'étude repose le plus souvent sur *Gene Ontology* [Ashburner 2000], une ontologie recouvrant les concepts associés aux composantes cellulaires, aux fonctions moléculaires et aux processus biologiques. La Fig. 23 donne un extrait des concepts associés au gène EPO (L'érythropoïétine (EPO) entraîne une augmentation du nombre de globules rouges dans le sang, cette protéine est très prisée par certains cyclistes !).

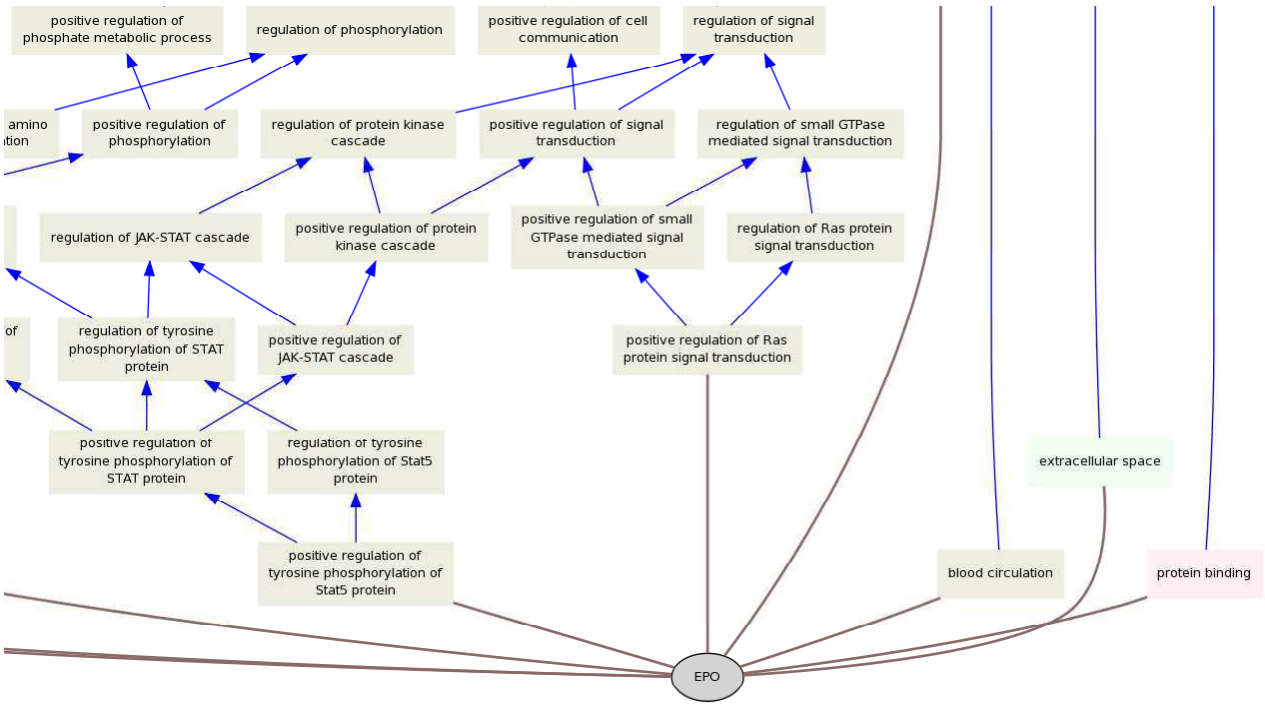


Fig. 23. *Extrait des concepts de Gene Ontology associés à l'EPO.*
La figure a été créée avec l'outil GenNav (http://mor.nlm.nih.gov/perl/gennav.pl).

4.4. Techniques de fouille de données pour le transcriptome

L'application des techniques de *data mining* aux données génomiques est considérée comme une tâche particulièrement ardue et représente en soi un véritable défi scientifique [Piatetsky-Shapiro 2003]. Elle repose sur une analyse exploratoire de données en recherchant de manière systématique des relations potentielles sans connaissances préalables sur la nature de ces relations [Wren et. al. 2005]. Une des applications les plus basiques est la recherche de patrons d'expression : dans l'étude du transcriptome, il s'agit de découvrir des gènes co-exprimés, qui sont des groupes de gènes dont l'expression est semblable sur un ensemble d'échantillons (cf. Fig. 21).

Un exemple typique d'outil de *data mining* est GenMiner [Martinez et al. 2008]. Ce logiciel est un système de découverte de règles d'associations dédié aux données génomiques. Il permet l'analyse de jeux provenant de sources multiples de données en prenant en compte des observations discrètes (comme les annotations) et continues (comme les données d'expression). D'autres auteurs ont cherché à croiser des données hétérogènes. En partant de données d'interaction protéique et de données d'expression, on peut ainsi renforcer la pertinence des résultats obtenus [Pei et al. 2007]. En utilisant plusieurs sources génomiques, il est possible d'analyser des gènes annotés dans des voies de signalisation métaboliques, des régulateurs de transcription et des catégories de *Gene Ontology* [Carmona-Saez 2006]. Les difficultés de ces approches sont essentiellement liées à l'interprétation et à la visualisation des informations extraites. Certains auteurs (notamment [Kennedy 2008]) ont recherché une méthodologie visant à intégrer les différentes techniques expérimentales dans un contexte biomédical incluant les aspects cliniques (données sur les patients), mais un système véritablement générique pour l'analyse de données génomiques hétérogènes reste encore à concevoir.

Outre la découverte de patrons, un but essentiel en biologie consiste en la construction de modèles capables de prédire le comportement de systèmes complexes. Un enjeu important est donc de créer

un réseau causal de gènes à partir des données expérimentales. Une approche basée sur des données d'expression, des sites de liaison aux facteurs de transcription, des interactions protéiques, a notamment permis de mettre en évidence le rôle à grande échelle des régulateurs dans l'expression des gènes [Zhu 2008].

Chapitre 5. Analyse statistique implicative sur les données de puces à ADN

L'étude du transcriptome représente un enjeu considérable, tant du point de vue de la compréhension des mécanismes du vivant que des applications cliniques et pharmacologiques. Malheureusement, les données d'expression sont entachées de multiples bruits. D'une part, la complexité du protocole expérimental entraîne inévitablement une imprécision des mesures d'expression. D'autre part, la variabilité naturelle de l'activité cellulaire induit des différences notables d'amplitude d'expression entre les gènes, phénomène que l'on peut apprécier en considérant plusieurs patients présentant le même phénotype.

L'analyse du transcriptome a pour objectif d'aider le biologiste à résoudre des questions essentielles pour la compréhension du fonctionnement des gènes :

- quel est le rôle d'un gène, dans quel processus biologique est-il impliqué ?
- comment les gènes sont-il régulés ?
- quelles sont les variations d'expression des gènes selon le type et l'état d'une cellule ?
- quels sont les gènes impliqués dans une pathologie particulière ?

La fouille de données se prête bien à ce type de problématique, étant donné le volume important des données et la complexité attendue des réponses recherchées. Parmi les techniques à notre disposition, les règles d'association semblent particulièrement bien adaptées : elles établissent des relations entre des groupes d'attributs et les questions précédentes peuvent aisément se formaliser en terme d'association entre des gènes, des maladies ou des fonctions. L'analyse implicative est le domaine de recherche qui nous a particulièrement intéressés, dans la mesure où cette thématique repose sur une base statistique.

Nous proposons une méthode d'analyse implicative des règles d'association sur les données du transcriptome. Elle utilise l'approche développée initialement par Régis Gras [Gras 1996] en considérant non pas les mesures elles-mêmes, mais le rang des observations. Cette optique permet de s'affranchir des valeurs numériques en considérant des zones de classement dans les mesures d'expression. S'intéresser au classement a l'avantage d'améliorer la robustesse des algorithmes en les rendant insensibles à des transformations monotones des données. Le traitement des données incertaines ou imprécises a déjà fait l'objet de travaux dans le cadre de l'analyse implicative. Dans [Gras 2001], la méthode proposée consiste à définir une partition optimale des données puis à rechercher des implications entre les intervalles ainsi constitués. Une approche parallèle a été développée, basée sur la logique floue [Gras 2005]. Ces deux méthodes reposent sur une partition préalable des mesures avant l'analyse implicative proprement dite. Comme les distributions de nos données se sont avérées monomodales, nous avons préféré rechercher directement l'implication

optimale sans passer par un prétraitement des données qui risquerait d'introduire un biais dans l'extraction des règles.

5.1. Règles d'association et intensité d'implication

La recherche de règles d'association est une problématique de fouille de données consistant à découvrir des relations entre plusieurs variables. Dans sa version la plus commune, une règle d'association se caractérise de la manière suivante : soit $A=\{a_1,a_2,...,a_m\}$ un ensemble d'attributs binaires et $E=\{i_1,i_2,...,i_n\}$ un ensemble d'individus. A chaque individu est associé un sous-ensemble de A. Une règle d'association est définie par une implication de la forme $X \rightarrow Y$, où $X,Y \subseteq A$ et $X \cap Y=\emptyset$. L'ensemble X est la prémisse de la règle et Y sa conclusion. La règle formalise une association entre les items de X et ceux de Y, énonçant le fait qu'un individu ayant les attributs de X aura tendance à également posséder ceux de Y. Une règle d'association n'est donc pas une règle logique mais une règle statistique admettant généralement quelques contre-exemples si la tendance implicative n'est pas remise en cause.

Un exemple classique de règle d'association est celui du « panier de la ménagère » où sont recherchées les règles d'achat de clients de supermarchés [Agrawal 1993]. Les règles d'association ont également été utilisées en bio-informatique. Dans [Carmona-Saez 2006], les auteurs appliquent la découverte de règles d'association pour l'analyse intégrative de données de puces à ADN. Les données traitées comprennent des données d'expression ainsi que des annotations de gènes. Les règles extraites expriment notamment des relations entre des catégories (classes de gènes prédéfinies) et des observations expérimentales.

Par exemple, la règle `cystein metabolism → T7 [sous-expression]` indique qu'un gène impliqué dans le métabolisme de la cystéine possède généralement une sous-expression observée au temps 7 dans une étude temporelle de puces à ADN sur *Saccharomyces cerevisiae* (levure du boulanger). Deux indicateurs sont généralement utilisés pour l'étude des règles d'association. Le premier indicateur sert à estimer l'importance de la règle en termes d'individus concernés. Cet estimateur est appelé le support, défini comme la fréquence des individus respectant une règle. Le deuxième type de mesure permet d'estimer la qualité de la règle. Un exemple typique de mesure de qualité est la confiance de la règle, définie comme le rapport du nombre d'individus vérifiant la règle sur le nombre d'individus vérifiant la prémisse de la règle (estimation de la probabilité conditionnelle). L'intensité d'implication [Gras 1996] est un autre exemple de mesure de la qualité d'une règle. L'intensité d'implication $\varphi(A \rightarrow B)$ mesure la qualité d'une règle $A \rightarrow B$ où α et β représentent les ensembles d'attributs vérifiant respectivement la prémisse et la conclusion de la règle. Cette mesure évalue l'étonnement statistique d'observer si peu de contre-exemples dans une règle d'association sous l'hypothèse d'indépendance des ensembles α et β.

Soit *E*, l'ensemble des individus observés. Soit X et Y un tirage aléatoire de deux parties de *E* choisies indépendamment, de mêmes cardinaux respectifs que ceux de α et de β. Le nombre de contre-exemples est défini par $card(X \cap \neg Y)$ (le cardinal de l'intersection de X avec le complémentaire de Y dans *E*). La variable $card(X \cap \neg Y)$ est une variable aléatoire dont $card(\alpha \cap \neg \beta)$ est une valeur observée. La mesure de qualité repose sur la probabilité d'observer $card(\alpha \cap \neg \beta)$ à partir des tirages aléatoires de X et Y. Plusieurs hypothèses de tirage sont envisageables : elles peuvent se baser sur des lois hypergéométriques, binomiales ou de Poisson. Quelque soit le modèle utilisé, l'intensité d'implication est toujours définie par l'expression suivante :

$$\varphi(A \rightarrow B)=1-\text{Probabilité}(card(X \cap \neg Y) \leq card(\alpha \cap \neg \beta)) \qquad (16)$$

L'avantage de l'intensité d'implication par rapport à la confiance est que cette mesure tient compte des effectifs d'individus. Par exemple, une règle dont la conclusion est quasi-systématiquement vérifiée par les individus de *E* aura peu d'intérêt pour un décideur. Elle aura pourtant une confiance proche de 1 (règle jugée à tort comme intéressante) alors que l'intensité d'implication sera comme on le souhaite proche de 0 (règle de faible qualité).

5.2. Le concept d'intervalle de rang

Les données sont modélisées par une matrice *M(k, o)* de données numériques, où *k* représente un individu (gène), *o* une observation (échantillon) et *M(k, o)* la mesure d'expression effectuée. On note *m* le nombre d'individus et *n* le nombre d'observations. Nous appelons profil d'un individu *k* le vecteur $pr(k)=(M(k,o), o\in[1,n])$. Le profil d'expression d'un gène est donc l'ensemble des valeurs d'expression pour ce gène, mesurées sur l'ensemble des observations (le profil correspond à une ligne de la matrice d'expression).

La Table 9 indique un extrait de la matrice concernant un gène k.

gène	observation			
	1	2	3	4
k	56	32	127	12

Table 9. *Exemple de profil d'expression sur 4 observations.*
Les observations o sont numérotées de 1 à n. Les valeurs d'expression correspondent à la mesure de l'expression du gène k dans l'observation o.

Nous supposons l'existence d'un opérateur *rang* qui délivre les observations d'un profil dans l'ordre croissant des valeurs d'expression. Plus précisément, nous avons $rang(pr(k))=(o_1,o_2,\cdots,o_n)$, où les observations $o_i\in[1,n]$ vérifient la condition suivante : $M(k,o_1)\leq M(k,o_2)\leq\cdots\leq M(k,o_n)$.

Pour le profil décrit en Table 9, l'opérateur de rang fournit la liste ordonnée suivante : $(o_1=4, o_2=2, o_3=1, o_4=3)$. La composante $o_1=4$ désigne le fait que la valeur d'expression la plus faible a été mesurée sur l'observation 4 (valeur d'expression de 14). De même, la valeur la plus forte correspond à l'observation $o_4=3$ (valeur d'expression de 127).

Nous nous proposons d'extraire des règles d'association entre profils d'expression. Ces règles vont mettre en évidence des intervalles d'étude. Un intervalle de rang permet de déterminer des niveaux de classement sans avoir à définir des seuils de mesure. Pour l'analyse du transcriptome, ces intervalles se rapportent à des niveaux d'expression, telles que la sur-expression ou la sous-expression. Notre approche considère tous les intervalles possibles, à savoir l'ensemble des sous-intervalles de [1, *n*] :

$$I=\{\,[p,q], (p,q)\in[1,n]^2,\ p\leq q\,\}$$

A un intervalle $i\in I$ est associé un *intervalle de rang* $r_k(i)$, défini comme suit : $r_k(i)=(o_j\in rang(p(k))), j\in i$. Comme les valeurs sont ordonnées par l'opérateur *rang*, le concept d'intervalle de rang renvoie aux observations relatives à des niveaux d'expression. Par exemple, un intervalle de [1,*x*] définit la zone la plus basse de classement (les *x* valeurs d'expression les plus faibles d'un profil d'expression). On peut faire l'analogie avec des notes d'étudiants dans une matière *k* : l'intervalle *i*=[1,10] définit le bas de classement dans cette matière et l'intervalle de rang $r_k(i)$ désigne les étudiants qui ont obtenu les dix plus mauvaises notes.

La Table 10 donne des exemples d'intervalles de rang pour des niveaux d'expression.

	niveau d'expression		
	bas	moyen	haut
intervalle	[1,2]	[2,3]	[3,4]
intervalle de rang	(4,2)	(2,1)	(1,3)

Table 10. *Intervalles de rang.*
Dans cet exemple, le niveau d'expression moyen définit le milieu de classement (intervalle [2,3] excluant la valeur la plus faible (indice 1) et la plus forte (indice 4). Cette zone moyenne dans le classement correspond aux observations 2 et 1. La Table 9 indique les valeurs correspondantes d'expression (respectivement 32 et 56).

L'analyse statistique implicative nous permet d'étudier la relation entre deux intervalles de rang $r_A(i)$ et $r_B(j)$. Une règle de type $r_A(i) \rightarrow r_B(j)$ peut s'interpréter de la manière suivante :

- l'intervalle *i* détermine un certain niveau d'expression pour un gène A,
- l'intervalle *j* détermine un certain niveau d'expression pour un gène B,
- la prémisse et la conclusion définissent les observations relatives à ces niveaux d'expression,
- la règle exprime le fait qu'une observation appartenant à la prémisse se retrouve généralement dans la conclusion.

De manière plus concrète, ce type de règles permet d'exprimer des relations entre les gènes. Par exemple, une sous-expression d'un gène A implique une sur-expression d'un gène B, le support de la règle étant un groupe d'observations.

Comme il a été vu en début de chapitre, l'intensité d'implication $\varphi(r_A(i),r_B(j))$ mesure la qualité d'une telle règle. Notons que dans notre approche l'hypothèse nulle d'indépendance revient à affirmer que l'opérateur de classement *rang* n'apporte aucune information utile. Dans ce cas, prendre deux intervalles de rang reviendrait à sélectionner les observations au hasard, sans tenir compte du classement des valeurs d'expression.

Nous étendons la définition (16) de l'intensité d'implication par l'expression suivante :

$$\varphi(A \rightarrow B) = \max\left\{\varphi(r_A(i),r_B(j)), (i,j) \in I^2\right\} \qquad (17)$$

Dans cette formulation, la qualité de la règle liant un gène A à un gène B est déterminée par la relation la plus forte entre des intervalles de rang $r_A(i)$ et $r_B(j)$ pour tout intervalle de *I*.
L'intérêt de cette définition est de considérer les relations remarquables entre gènes, sans définir arbitrairement des intervalles prédéterminés.

5.3. Intérêt de l'approche implicative pour l'étude des données d'expression

En matière d'étude du transcriptome, les analyses les plus couramment menées par les biologistes sont basées sur des mesures de corrélation entre profils d'expression. Ces mesures présentent l'inconvénient d'être globales : elles font intervenir l'ensemble des observations, alors que la mesure implicative recherche une correspondance optimale entre des sous-ensembles d'observations. Pour expliciter la différence entre ces deux approches, nous allons considérer un exemple de règles d'association entre deux gènes. Ces gènes appartiennent au génome de l'espèce Saccharomyces cerevisiae, communément dénommée levure du boulanger. Nous avons repris les

données de puces sélectionnées dans [Gasch 2002]. Nous avons retenu 89 conditions expérimentales différentes correspondant à différents stress induits tels que le choc thermique.

La Fig. 24 représente les profils d'expression des deux gènes impliqués dans la règle suivante : CHA1 → SAM1. Le gène CHA1, intervenant dans le catabolisme de la thréonine, est clairement sous-exprimé en réponse à un signal de déficience en acide aminé, et dans une moindre mesure, en nitrogène. SAM1, un gène interférant dans le métabolisme de la méthionine, est sur-exprimé pour le même jeu d'observations. Ce jeu correspond à environ 9 % des conditions.

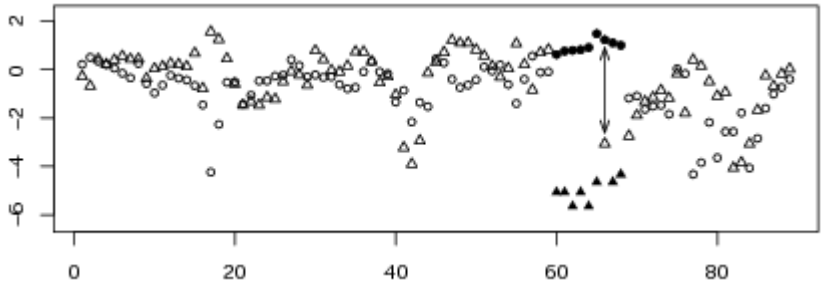


Fig. 24. *Profils des gènes CHA1 et SAM1.*
L'axe des abscisses représente 89 conditions expérimentales. L'axe des ordonnées représente les mesures d'expression. Le profil du gène CHA1 (YCL064C) est représenté par des triangles et celui de SAM1(YLR180W) par des cercles. Les figures pleines correspondent aux observations appartenant aux intervalles de rang qui maximisent la valeur de l'intensité d'implication. On remarque que ces observations sont communes à une exception près, indiquée par la double flèche : l'observation de SAM1 qui est moins sous-exprimée, n'apparaît pas dans le groupe alors qu'il est bien présent dans celui de CHA1.

Comme le montre la Table 11, les indices usuels de corrélation ne peuvent déceler de telles associations. Les valeurs obtenues sont trop faibles pour être retenues dans une analyse alors que la mesure implicative exprime que le risque de rencontrer un faux-positif (association de même nature relevée au hasard) est inférieur à un pour mille.

Méthode	Valeur
Intensité d'implication	0,9992
Indice de corrélation de Pearson	0,16
Indice de corrélation de Kendall	0,0089

Table 11. *Comparaison de mesures entre profils d'expression.*

5.4. Une définition implicative du concept de gène discriminant

Les données de puces concernent généralement un nombre important de gènes. La plupart d'entre eux n'aident pas à discriminer les classes, soit parce que leurs expressions ont une faible amplitude de variation, soit parce que leur expression est indépendante de l'ensemble *C* des classes d'observations. Une étape de sélection des gènes est donc nécessaire. La technique la plus communément partagée repose sur la puissance discriminative des tests statistiques tels que le test de Student ou ANOVA (voir [Chen 2005] pour une synthèse des mesures utilisées sur les données d'expression). Nous proposons d'utiliser l'intensité d'implication pour déterminer les gènes les plus discriminants. On appellera fonction d'étiquetage la fonction $L(o_k) = c_l$, où o_k est une observation appartenant à l'ensemble *O* des observations et c_i une classe de *C*. Notre approche est fondée sur des règles d'association de la forme : $r_g(i) \rightarrow O_c$. Cette règle comporte les deux champs suivants :

- la prémisse désignant l'intervalle de rang associé au gène *g* et à l'intervalle *i,* soit l'ensemble des observations situées dans cet intervalle pour le gène considéré,
- la conclusion désignant les observations de classe *c.*

La règle compare donc deux ensembles d'observations et détermine la surprise statistique de rencontrer autant d'individus de classe *c* dans la conclusion. Cette règle peut donc être explicitée de la façon suivante : si pour une observation *o* donnée, la mesure d'expression relative au gène *g* figure dans la zone de classement défini par l'intervalle *i,* alors cette observation *o* appartient à la classe *c* (avec un niveau de confiance donné par la mesure implicative). Nous restreignons par la suite le domaine des intervalles en imposant des niveaux d'expression ayant une signification biologique précise, à savoir la sous-expression (intervalle de la forme [1, q]) ou la surexpression (intervalle [p, n]). L'ensemble I des intervalles est donc désormais : $I=\{[p,q],(p,q)\in[1,n]^2,\ p\leq q\ ,p=1\wedge q=n\}$.

La règle $r_g(i)\rightarrow O_c$ peut dès lors s'interpréter plus simplement : elle signifie que si on observe une expression remarquable (faible ou forte) du gène *g* sur un patient *o*, alors ce patient possède probablement le phénotype *c*. Cette implication correspond bien à la notion de gène discriminant. Le pouvoir discriminant d'un gène g vis à vis d'une classe c est exprimée par l'expression suivante :

$$\varphi_C(g)=\max(\ \varphi(\ r_g(i),O_c),\ i\in I,\ O_c=\{o,L(o)=c\}) \qquad (18)$$

La mesure $\varphi_c(g)$ est définie par la maximisation de l'intensité d'implication des règles de classification de type $r_g(i)\rightarrow O_c$. Une remarquable propriété de l'intensité d'implication est de prendre en compte le nombre d'observations appartenant à O_c: pour un intervalle fixe *i*, le fait d'accroître la population O_c entraîne une diminution de la qualité de la règle. Il est en effet normal de retrouver une proportion notable d'observations de classe *c* si cette dernière est surreprésentée.
La mesure $\varphi_c(g)$ permet d'opérer une sélection des gènes discriminants en classant les gènes par ordre décroissant des intensités d'implication et en ne retenant que les k gènes les plus discriminants.

Cette technique a été utilisée sur une étude portant sur deux types de leucémie [Golub 1999]. La leucémie se caractérise par une prolifération maligne de cellules d'origines hématopoïétiques peu matures et rapidement diffusantes. Cette maladie se caractérise par une atteinte massive de la moelle osseuse, due au développement de lymphomes malins. On distingue les leucémies aiguës lymphoblastiques (notées ALL par la suite pour *Acute Lymphoblasic Leukemia*) des leucémies aiguës myéloblastiques (notées AML pour *Acute Myeloid Leukemia*). La distinction entre ces deux formes est essentielle pour le succès des thérapies envisagées : le traitement diffère selon l'une ou l'autre de ces deux classes de leucémie. Le jeu de données comporte 38 patients (27 patients ALL et 11 patients AML) et concerne 3571 gènes.

Nous nous proposons d'extraire dans ce jeu les gènes les plus discriminants en appliquant l'algorithme de sélection décrit précédemment. Nous avons étudié le pouvoir discriminant des gènes du jeu complet sur la base de la mesure implicative, dont nous donnons ici une version logarithmique : $\lambda_c(g)=-\log_{10}(1-\varphi_c(g))$. Cette transformation rend les valeurs plus facilement interprétables en donnant des ordres de grandeur. Une règle de qualité $\lambda_c(g)$ possède ainsi un risque de $10^{-\lambda_c(g)}$ d'être dû au hasard. L'analyse révèle que 10% des gènes ont un pouvoir discriminant important ($\lambda_c(g) > 3{,}5$ soit $\varphi_c(g) > 0{,}9997$).

Bien que 300 gènes puissent être considérés comme discriminants, les auteurs de l'étude n'ont retenu qu'une liste de 50 gènes les plus informatifs, tout en indiquant le caractère arbitraire de ce

nombre. Comme ils ont développé une technique originale de sélection, il est intéressant de comparer leur approche avec la nôtre. Nous avons recherché les gènes les plus discriminants pour les classes ALL et AML (50 gènes ou total, en retenant les 25 gènes ayant la mesure implicative logarithmique la plus forte pour chaque classe). Nous avons obtenu une liste de gènes discriminants dans laquelle figurent 14 gènes appartenant à la liste publiée par les auteurs. La Table 12 compare le pouvoir discriminant des deux jeux. On remarque que le gène le plus discriminant est commun aux deux jeux de gènes et que la moyenne est du même ordre de grandeur. On observe cependant une plus grande dispersion dans le jeu de Golub *et al*. De même, la valeur médiane est significativement plus faible par rapport à notre liste.

Liste de gènes	max	médiane	min	moyenne	variance
Golub et al.	9.08	5.44	4.55	5.12	10.16
notre liste	9.08	6.47	5.49	5.92	11.80

Table 12. *Pouvoir discriminant des gènes sélectionnés.*
Comparaison des valeurs de $\lambda(g)$ *sur les 50 gènes sélectionnés par notre méthode et celle de Golub et al (1999).*

Pour comparer la puissance discriminative de ces deux jeux de gènes, nous avons procédé à une validation croisée sur les données en utilisant le même classifieur, à savoir les 3 plus proches voisins. Cette technique a été retenue parce qu'elle prend en considération la distance entre points dans tout l'espace des gènes, contrairement à d'autres méthodes qui vont privilégier certains gènes (arbres de décision, forêt aléatoire, séparateurs à vastes marges) et qui apporteraient donc un biais pour la comparaison. Nous avons appliqué l'algorithme sur la matrice d'expression réduite comportant les mêmes patients et portant sur le jeu de gènes considéré. La Table 13 montre que notre sélection donne des résultats supérieurs à celle des auteurs.

% test	Golub & al.	Nos résultats
50%	3.11	0.79
25%	1.67	0.11
10%	1.00	0.00
2.6%	0.00	0.00

Table 13. *Comparaison des taux d'erreurs en validation croisée.*
La première colonne indique le pourcentage considéré en test sur le jeu de données. Les deux autres colonnes donnent les taux d'erreurs en pourcentage sur 100 jeux de validation aléatoires.

5.5. Application à la classification de tumeurs

La section précédente a permis de définir une mesure du pouvoir discriminant des gènes. Comme nous avons basé notre algorithme de sélection sur des règles de classification de type $r_g(i) \rightarrow O_c$, il est naturel d'envisager d'utiliser celles-ci pour prédire la classe d'un patient d'après son profil d'expression. Cette catégorisation est un des enjeux majeurs de la technologie des biopuces : elle permet de diagnostiquer l'existence d'un cancer à un stage précoce, lorsque la maladie s'exprime

dans les cellules sans qu'on observe encore des signes cliniques manifestes. Par ailleurs, la prédiction basée sur l'expression des gènes permet de distinguer entre différents types de tumeurs même si leur apparence morphologique tumorale est identique.

Pour caractériser la capacité prédictive des règles extraites, nous proposons une approche supervisée comprenant un jeu d'apprentissage $\mathcal{A} = \{ \mathcal{G}, M, \mathcal{O}, \mathcal{L}, C\}$, où $\mathcal{G}$ est un ensemble de gènes, M les mesures effectuées sur un ensemble $\mathcal{O}$ d'observations, et $\mathcal{L}$ la fonction qui attribue à chaque observation une classe de C. Le principe de la prédiction d'un échantillon à partir d'une expérimentation peut être décrit comme suit : soit o une observation nouvelle de classe inconnue sur laquelle a été effectuée une mesure d'expression sur l'ensemble des gènes $\mathcal{G}$ et soit $pr(o)$ le profil d'expression correspondant à ces mesures, il s'agit de relever les prémisses des règles de classification que respecte o et définir $\mathcal{L}(o)$ comme étant la classe la plus souvent rencontrée en conclusion de ces règles. On ne peut en pratique vérifier directement si o respecte la prémisse P d'une règle $r_g(i) \rightarrow O_c$, où O_c désigne les observations de classe c. En effet, $r_g(i)$ est défini à partir de l'ensemble $O \in \mathcal{A}$, ensemble d'apprentissage dans lequel ne figure pas l'observation o. Pour pallier cette difficulté, on vérifie si la nouvelle observation s'insère entre les rangs du classement effectué au moment de l'apprentissage. La pratique opératoire est donc la suivante : soit $s(P)=\min\{M[g,o]; o \in P\}$ et $S(P)=\max\{M[g,o]; o \in P\}$, on dira que o respecte la prémisse P si la condition suivante est réalisée : $s(P) \leq M[g,o] \leq S(P)$.

La Table 14 montre les jeux de données publiques utilisées pour tester la pertinence de notre approche.

Jeu de données	*Publication*	*# tissus*	*# classes*	*# gènes*	*type d'étude*
Cerveau	[Pomeroy 2002]	42	5	5 597	sous-types tumoraux
Colon	[Alon 1999]	62	2	2 000	tissus sains / malades
Leucémie	[Golub 1999]	72	2	3 571	sous-types tumoraux

Table 14. *Jeux de données publiques sur le cancer.*

La Table 15 présente les résultats obtenus selon la technique du *leave-one-out* (validation croisée en prenant l'ensemble des observations en apprentissage moins une, cette dernière servant de jeu de test, ce principe étant répété pour chaque observation). Sur les trois jeux de données décrits précédemment, notre méthode atteint des performances comparables aux autres classifieurs. Ce résultat est d'autant plus remarquable que notre algorithme est relativement frustre, puisqu'il s'agit d'un simple comptage de règles de classification.

méthode	*cerveau*	*colon*	*leucémie*
Règles de classification	**14.3**	**12.9**	**2.8**
Gene clustering [Dettling 2002]	11.9	16.1	2.8
Fuzzy cmeans [Wang 2003]	14.3	11.4	4.1
Forêt aléatoire [Diaz-Uriarte 2006]	19.0	14.5	2.8
Séparateurs à vastes marges [Lee 2003]	11.9	12.9	2.8
3 plus proches voisins [Yeang 2001].	23.8	22.6	1.4

Table 15. *Comparaison des méthodes de classification.*
Le tableau indique les taux d'erreurs selon la technique du leave-one-out. Bien évidemment, la méthode basée sur les règles de classification effectue à chaque test une extraction de règles sur le jeu d'apprentissage (le candidat testé ayant été retiré de ce jeu).

5.6. Conclusion

L'approche implicative, appliquée aux données d'expression, présente plusieurs avantages. En premier lieu, elle est plus fine que les techniques qui mesurent des relations de similarité globale. Par mesure globale, nous entendons des estimations basées sur l'ensemble des observations. Il est clair que si un certain nombre d'observations ne participent pas à une relation, ces informations apportent un bruit qui masque l'association entre gènes, comme nous l'avons montré sur un exemple d'expérimentation sur la levure. De la même manière, une association entre conditions expérimentales peut être masquée par des gènes qui ne sont pas régulés de manière coordonnée avec un autre groupe de gènes. C'est la raison pour laquelle une analyse implicative est plus performante que des techniques basées sur des corrélations.

Un deuxième intérêt de notre approche est lié à la robustesse de l'analyse de rang. On remarquera que l'analyse de classement est invariante par rapport à toute transformation monotone des données. Cette propriété est particulièrement utile dans le cadre de données de puces qui subissent un grand nombre de prétraitements (transformation logarithmique, normalisation, …).
Enfin, on rappelle que l'implication est une information orientée, contrairement aux mesures de similarité qui sont symétriques. Cette propriété peut être exploitée dans le cadre du transcriptome. On sait en effet que les gènes sont activés par le biais de facteurs de transcription, qui sont eux-mêmes exprimés dans la cellule.

Nous avons proposé une méthode originale de sélection des gènes informatifs. La pertinence de la méthode a été vérifiée en démontrant le pouvoir prédictif du jeu sélectionné. La découverte de gènes discriminants est d'une grande importance pour les applications cliniques, car elle permet de définir des méthodes de diagnostic fiables et relativement peu coûteuses.

Nous avons développé un algorithme d'extraction de règles de classification. L'avantage d'une méthode de classification basée sur les règles est qu'elle délivre une information aisément interprétable par un expert biologiste, contrairement à des méthodes telles que les Séparateurs à Vastes Marges. Malgré sa simplicité relative de mise en œuvre, notre algorithme s'est révélé aussi performant que les techniques les plus éprouvées dans ce domaine.

Chapitre 6. Intégration et analyse d'identifiant de gènes

Les données de puces à ADN ne cessent de croître. *Gene Expression Omnibus* (GEO, http://www.ncbi.nlm.nih.gov/geo/), un site du *National Center for Biotechnology Information* (NCBI, appartenant au *National Institutes of Health*), collecte les données publiques sur le transcriptome [Edgar 2002]. GEO rassemble actuellement plus de 500 000 échantillons issus de plus de 20 000 études.
De telles bases renferment une quantité considérable d'informations utiles. L'hétérogénéité des technologies est cependant un obstacle majeur pour la poursuite d'études intégratives. Par exemple, les données de GEO ont été mesurées sur un ensemble de plus de 8 000 supports technologiques distincts.
La conception des puces diffère fortement selon les fournisseurs, chaque fournisseur offrant un vaste éventail de produits. Le nombre de sondes et la nature des sondes est spécifique à chaque support technologique. L'intégration de ces données est donc une problématique majeure pour les chercheurs. Ce chapitre décrit un effort d'intégration des identifiants de gènes pour faciliter la comparaison d'études du transcriptome. Ce travail comporte les deux aspects suivants :

- le premier aspect concerne la conception d'un outil d'intégration au service de la communauté des bio-informaticiens. Ce volet essentiellement technologique nécessite la conception d'une base de données pour faciliter la conversion des identifiants entre eux et une application pour traiter plus aisément des listes d'identifiants.
- le deuxième aspect présente des algorithmes originaux et des applications démontrant la pertinence d'une telle approche. Ce deuxième volet, d'essence plus scientifique, est une tentative pour s'abstraire des données d'expression et pour pratiquer des analyses à un niveau supérieur. Cette démarche repose sur l'analyse de listes d'identifiants en lieu et place des valeurs d'expression. Notons que cette approche reprend dans sa philosophie celle abordée dans le chapitre précédent : l'analyse statistique implicative portait sur des listes d'observations plutôt que sur les données d'expression.

6.1. Génomique intégrative

La génomique intégrative est sans aucun doute un axe stratégique de la recherche biomédicale. L'objectif de cette thématique est d'apporter une meilleure compréhension du comportement et de la fonction des gènes au sein d'un système biologique [Peng et al. 2007]. Un défi majeur en bio-informatique consiste à intégrer des informations disparates et à développer des études à grande

échelle pour analyser de multiples jeux expérimentaux [Bertone et al. 2001]. Il faut noter que le terme d'intégration ne signifie pas la simple collocation de différentes sources de données (à la manière par exemple de *GeneCards* [Rebhan 1997], une source qui centralise les liens web relatifs aux gènes), mais un concept garantissant l'unification de données hétérogènes dans un espace commun de recherche [Wren et al. 2005]. Malgré la maturation de nombreuses problématiques biologiques, force est de reconnaître qu'un outil unifié et générique d'analyse génomique n'est pas encore à l'ordre du jour.

L'intégration des données génomiques ne peut être envisagée sans la définition de standards pour assurer le partage de l'information et l'interopérabilité des outils [Field 2009]. Plusieurs efforts de standardisation sont en cours pour définir des descripteurs minimaux de données (projet *Minimum Information for Biological and Biomedical Investigations* (MIBBI), [Taylor 2008]), pour le développement coordonné d'ontologies (*Open Biomedical Ontologies* (OBO) *consortium*, [Smith 2007]) ou la représentation numérique des données biologiques (*Investigation/Study/Assay* (ISA) *infrastructure*, [Sansone 2007]).

Cet effort de normalisation a un coût qui n'est pas à négliger en termes de ressources matérielles et humaines. L'intégration de données génomiques hétérogènes obtenues à partir de technologies diverses peut cependant apporter une approche plus robuste dans les recherches biologiques. L'avantage premier de cette démarche est en effet une confiance accrue dans les résultats. Contrairement aux expérimentations basées sur une seule technique de génomique fonctionnelle, l'utilisation de plusieurs sources d'information permet d'accroître la spécificité et la sensibilité des travaux menés. La recherche des gènes responsables d'une pathologie passe le plus souvent par ce type d'étude croisée. Par exemple dans [Mootha 2003], les auteurs ont identifié le gène causant le syndrome de Leigh (encéphalomyopathie nécrosante) à partir de quatre jeux de données publiques d'expression. Ils ont défini un score de similarité des profils d'expression en se basant sur des gènes mitochondriaux connus. Une étude protéomique par spectrométrie de masse a parallèlement permis d'identifier les protéines présentes dans la mitochondrie. Le croisement de ces informations avec la région chromosomique associée (recherche des positions des gènes sur le génome mitochondrial) a conduit à isoler un seul gène candidat.

L'intégration de données a pour but d'associer des concepts distincts (ADN, ARN, protéines, pathologies, drogues,...). Dans les études menées jusqu'ici, la plupart des auteurs ont agrégé les informations de manière plus ou moins manuelle. De telles analyses sont fastidieuses et peuvent être entachées d'erreurs. Il est donc crucial de disposer d'outils adaptés pour permettre à l'utilisateur de développer de nouveaux types de requêtes [Ray 2003]. Un tel dispositif permettrait par exemple de sélectionner une région chromosomique, de définir une ou plusieurs pathologies, de préciser éventuellement des jeux de données d'expression. En retour l'utilisateur obtiendrait une liste de gènes d'intérêt.

L'objectif de ce chapitre est de construire une première brique dans cet ambitieux édifice que représente la génomique intégrative. La première étape consiste à normaliser les identifiants de base de la génomique fonctionnelle. Il est évident que l'identification correcte des gènes est une priorité pour établir des passerelles entre les diverses nomenclatures et synonymes existants. Pour s'en convaincre, il suffit de consulter les quelques résultats présentés en Fig. 26 : le gène TP53 est aussi connu sous le synonyme LFS1 et est référencé par une vaste collection d'identifiants provenant de multiples banques de données. Le biologiste doit donc faire face à une très grande hétérogénéité des nomenclatures pour désigner un gène. La multiplicité des désignations s'explique par différents facteurs. Le fabriquant d'une puce définira un identifiant pour une sonde déposée sur une puce commerciale ; une base de données biologiques établira sa propre nomenclature ; des

chercheurs auront désigné le même gène sous des appellations différentes. La plateforme de bio-informatique de Nantes a développé MADGene, une base de données accessible publiquement par internet qui permet de procéder à des conversions d'un référentiel en un autre ainsi que de rechercher des identifiants orthologues (en passant d'une espèce à une autre). Nous présentons cet outil avant de décrire des applications en analyse de données de puces à ADN.

6.2. La base de données MADGene

Le but de la base de données MADGene est de proposer une uniformisation des identifiants des gènes et d'établir des liens entre les différentes nomenclatures. Les références prises en compte proviennent des sources suivantes :

- Unigene
- HGNC
- Homologene
- EntrezGene
- Affymetrix
- GEO
- Ensembl

Les 17 espèces considérées sont les suivantes : *Anopheles gambiae, Arabidopsis thaliana , Bos taurus, Caenorhabditis elegans, Canis familiaris, Drosophila melanogaster, Danio rerio, Gallus gallus, Homo sapiens, Magnaporthe grisea, Mus musculus, Neurospora crassa, Oryza sativa, Pan troglodytes, Plasmodium falciparum, Rattus norvegicus, Saccharomyces cerevisiae*. La version actuelle de la base de données contient plus de 128 millions d'enregistrements.

La Fig. 25 donne un exemple d'interrogation de la base de données MADGene à partir du formulaire de requête disponible sur le site web *http://www.madtools.org/madgene*. Les résultats fournis par cet outil sont présentés en Fig. 26 (liste partielle). L'utilisateur peut consulter la liste en ligne ou la télécharger sous forme d'un fichier csv. La liste donnée en exemple concerne le gène p53 : un facteur de transcription régulant notamment l'apoptose (mort programmée de la cellule). Surnommé le gardien du génome, il est impliqué dans la moitié des cancers chez l'homme.

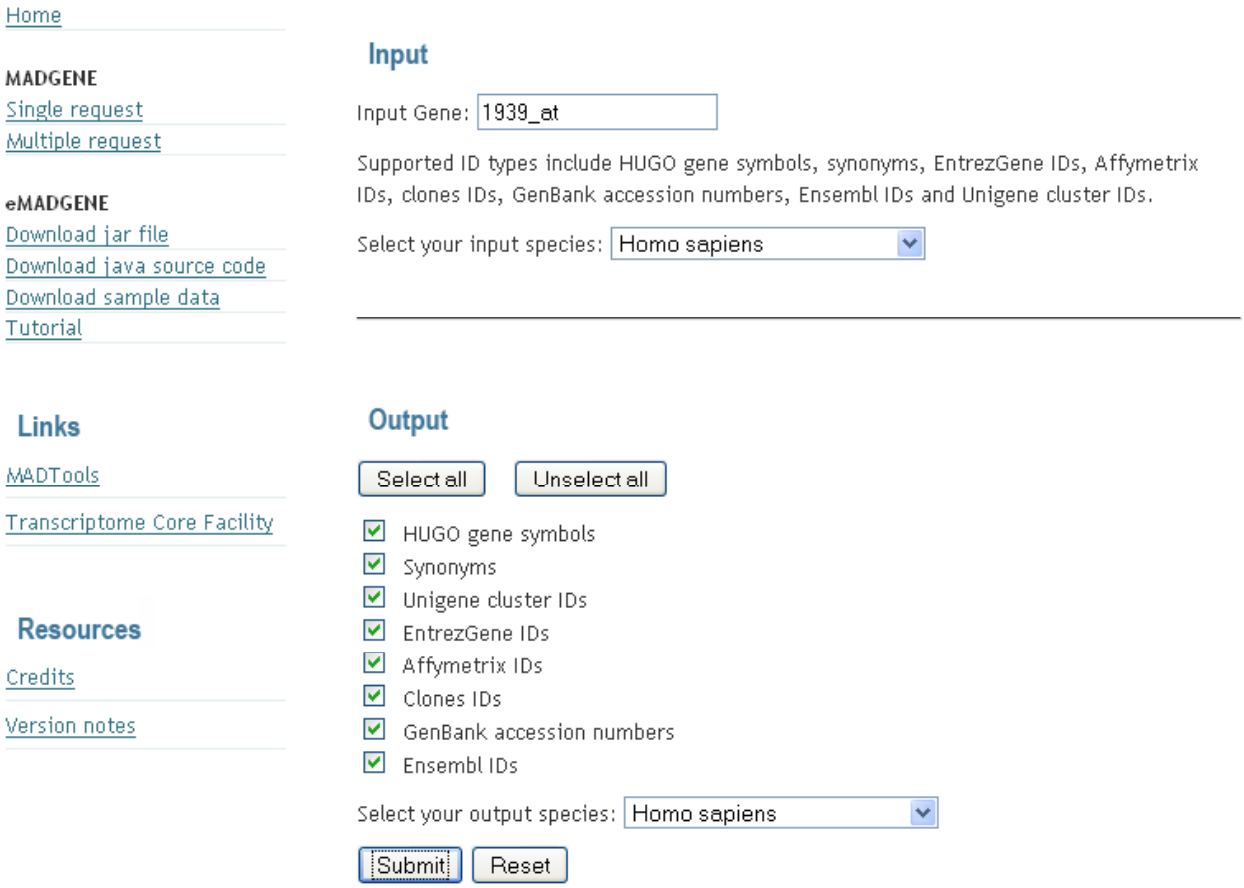


Fig. 25. *Interrogation de la base de données MADGene*
Une interface web accessible à l'adresse http://www.madtools.org/madgene permet d'interroger la base de données. La requête porte sur le code Affymetrix ; l'utilisateur peut définir les identifiants voulus et sélectionne l'organisme en entrée et en sortie.

Results for Hs

Symbol : TP53 Tumor protein p53

Synonyms:
p53
LFS1

Unigene cluster IDs: Hs.654481

EntrezGene IDs: 7157

Clone IDs:
IMAGE:5587461
IMAGE:5922321
IMAGE:5434546
IMAGE:4587646
IMAGE:3956624

Accession Numbers:
NM_001126114
DQ186648
NM_001126113

Accession Numbers:
NM_001126114
DQ186648
NM_001126113
DQ186649

Affymetrix IDs:
1939_at
1974_s_at
201746_at
211300_s_at
31618_at
201746_PM_at
211300_PM_s_at
M22898_at

ENSEMBL IDs:
ENSG00000141510
ENST00000414315
ENST00000359597
ENST00000420246
ENST00000419024
ENST00000413465

Fig. 26. *Résultat d'une requête sur la base de données MADGene*

La base de données MADGene est un moyen performant d'établir des liens entre des données de puces hétérogènes. Cependant, la mise en relation et l'analyse de ces informations reste une activité malaisée et laborieuse. Le biologiste ressent de plus en plus la nécessité de se confronter à des données non plus primaires, mais dérivées de ces travaux, telles que des listes de gènes discriminants ou des gènes présents dans des clusters issus de jeux expérimentaux indépendants. Ce besoin nous a poussés à concevoir eMADGene, une application Java dédiée à la manipulation de listes d'identifiants de gènes. Son but est de permettre à un utilisateur de MADGene de pleinement exploiter ses résultats.
Plusieurs outils existent pour la conversion des identifiants (Alibes et al. 2007; Bussey et al. 2003; Chen et al. 2007; Diehn et al. 2003; Huang et al. 2008). Le système intégré que nous proposons possède plusieurs avantages : la possibilité d'annoter des listes d'identifiants hétérogènes, la prise en compte d'un large spectre d'espèces et la possibilité d'analyser des listes multiples d'identifiants à travers un outil comportant des opérateurs relationnels ainsi que des tests statistiques variés.
L'application eMADGene comporte 23 opérateurs pour le traitement de listes, allant de simples opérateurs relationnels jusqu'à des méthodes statistiques élaborées. Ces fonctionnalités ont pour vocation :

- d'offrir des opérateurs relationnels pour manipuler des listes d'identifiants. eMADGene permet de transformer des données d'expression en liste d'identifiants, de réaliser entre autre des intersections, des unions ou des différences ensemblistes sur des listes de gènes.
- d'apporter des outils statistiques pour la découverte de biais de représentation dans des listes d'identifiants. Certains de ces outils sont basés sur le cardinal d'ensemble de gènes, d'autres exploitent le caractère ordonné de ces listes.

L'outil se présente sous la forme d'une application Java téléchargeable depuis le site des MADtools. Une interface graphique simple permet à l'utilisateur non programmeur de définir une séquence complexe d'opérations, comme le montre l'exemple de la Fig. 27.

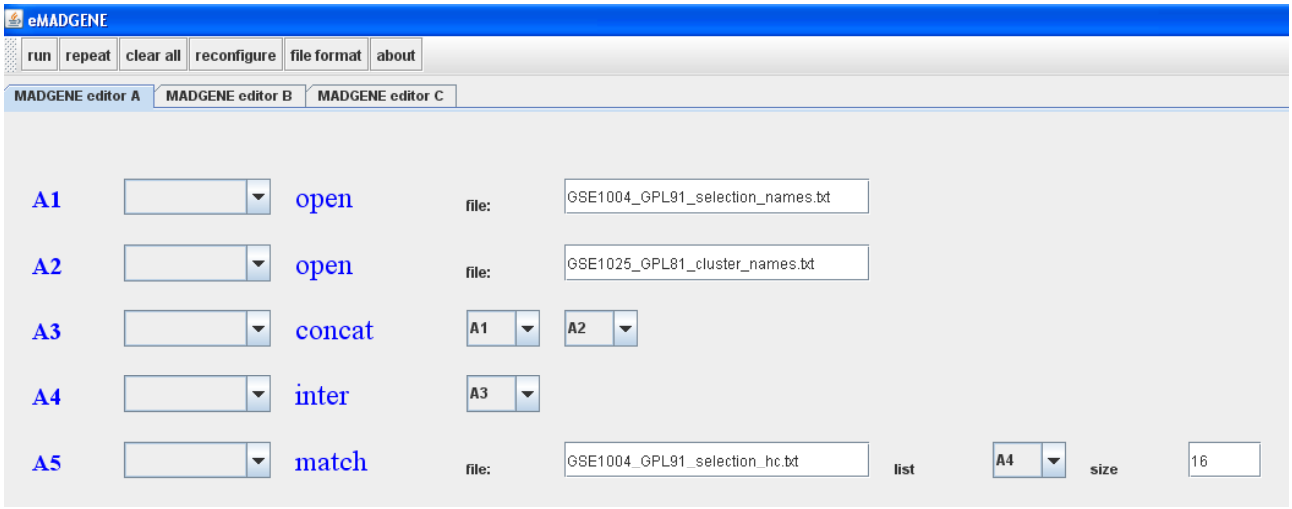


Fig. 27. *Analyse de listes de gènes avec eMADGene*
Les opérations peuvent être sélectionnées par une liste de choix. Les instructions sont codées par des identifiants (A1 à A5 sur l'exemple) et exécutées dans leur ordre de présentation. Les deux premières instructions ouvrent deux listes (liste des gènes d'une étude GSE 1004 (GPL 91) ainsi qu'un cluster de gènes extrait de l'étude GSE 1025 (GPL 81)). L'instruction A3 concatène les deux listes en une liste comportant deux colonnes. L'instruction suivante établit l'intersection des colonnes de la liste résultant de l'opération A3. La dernière instruction recherche le meilleur recouvrement entre les gènes communs (intersection de A1 et A2) et la liste des gènes de l'étude GSE 1004 (GPL 91) ordonnée par l'algorithme de classification hiérarchique d'Eisen (en utilisant une fenêtre glissante de taille 16). Une liste de cinq gènes a été retrouvée avec une p-valeur de 0.003.

Nous donnons deux applications pour illustrer la pertinence de cette approche. La première concerne une méta-analyse sur la maladie de Duchenne. Le terme de méta-analyse désigne une étude portant sur plusieurs expérimentations indépendantes [Moreau 2003]. Cette technique est largement utilisée en médecine pour améliorer la robustesse statistique en augmentant le nombre de cas observés. La deuxième application est une mise en correspondance de régions communes entre deux études indépendantes à partir de leur classification hiérarchique ascendante. Les jeux de données de ces deux illustrations sont extraits de la base de données *Gene Expression Omnibus*.

6.3. Application de l'outil eMADGene à la comparaison d'échantillons

Cette application met en valeur l'utilité d'une fonction précise du logiciel eMADGene : cet outil transforme des données d'expression en des listes de gènes ordonnés dans l'ordre croissant de ces valeurs. Chaque liste correspond dans ce cas à un échantillon. Il est dès lors possible d'étudier des biais de classement. Par exemple, on peut mettre en évidence des gènes présentant une surexpression dans certains échantillons (ces gènes se retrouvent préférentiellement en bas de classement). On présente un cas d'application relatif à une méta-analyse sur la maladie de Duchenne (*Duchenne muscular dystrophy* (*DMD*)). La maladie DMD est une forme génétique sévère de myopathie liée au manque d'une protéine essentielle (la dystrophine), une pathologie qui se caractérise par une irréversible dégénérescence du tissu musculaire. Cette illustration montre la robustesse de la méthode : des échantillons issus d'études indépendantes et basées sur des plateformes technologiques différentes se regroupent très nettement selon le type de tissu (sain ou DMD). L'avantage des études de rangs est d'être moins sensible au bruit de mesure, comme il en a été fait mention dans le chapitre précédent. Cette méthode permet donc de s'affranchir des problèmes de normalisation de mesures inter-études.

Nous avons procédé à une méta-analyse portant sur quatre séries provenant de la base GEO et portant sur la maladie de Duchenne. Toutes les expérimentations sont basées sur deux classes d'échantillons :

- les tissus sains (NHM pour *Normal Human Muscle*)
- les tissus malades (DMD pour *Duchenne Muscular Dystrophy*).

Les études sont basées sur sept plateformes GEO différentes (GPL 91 à GPL 97) et représentent au total 193 échantillons (96 NHM et 97 DMD). Tous les jeux ont été prétraités : filtrage, normalisation LOWESS et centrage médian des profils d'expression. Ce dernier traitement est important : il définit la valeur relative d'expression par rapport une référence propre à l'étude considérée (la médiane des valeurs prises sur les échantillons de l'étude). Comme les jeux sont pour la plupart équilibrés en terme de représentation des deux classes, les gènes ayant une incidence (directe ou indirecte) sur la pathologie auront un profil avec une expression en dessous ou au dessous de la médiane selon le type d'expression (activation ou inhibition). La Fig. 28 montre deux profils-types de gènes différentiels. Notons que la position relative d'un gène par rapport à la médiane est la même si l'on considère les valeurs ou les rangs.

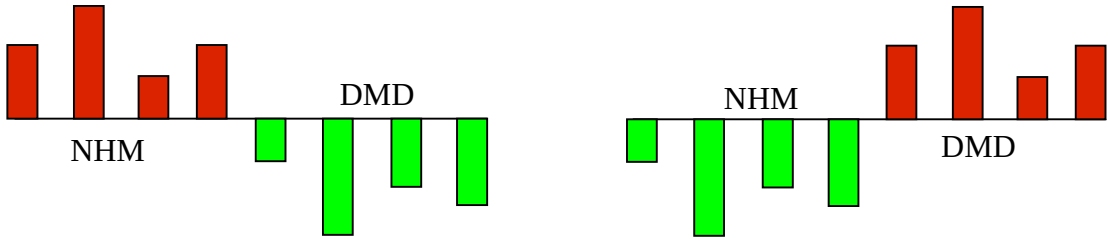


Fig. 28. *Profils de gène différentiel.*
Un gène possède une expression différentielle relativement à deux classes si le niveau d'expression est sensiblement différent d'une classe à l'autre (les deux classes sont symbolisées par des couleurs rouges et vertes). La figure de gauche présente une surexpression dans la classe NHM tandis que celle de droite présente une surexpression dans la classe DMD.

L'analyse que nous proposons repose sur une comparaison de l'expression des gènes sur chaque paire d'échantillons, à l'intérieur d'une étude ou entre deux échantillons issus de deux études différentes. Une telle comparaison pose en principe plusieurs problèmes :

- La première difficulté provient des conditions expérimentales hétérogènes qui ne permettent pas de comparer des valeurs numériques d'expression extraits de jeux différents. Le logiciel eMADGene transforme un échantillon de valeurs d'expression en une liste ordonnée d'identifiants classés par ordre croissant d'expression. Ce mode de représentation permet de s'affranchir des problèmes de normalisation inter-études. L'ordre des gènes exprime leur niveau d'expression : un gène situé dans la première moitié du classement sera sous-exprimé ; un gène situé dans la deuxième moitié sera surexprimé.
- Le deuxième problème concerne le fait que les plateformes ne mesurent pas les mêmes sondes et que ces derniers ne correspondent pas à un ensemble commun de gènes. La base de données MADGene permet d'unifier ces données en convertissant les identifiants des sondes en un référentiel commun (nous avons opté dans notre étude pour les symboles HUGO).

Pour valider notre approche nous avons comparé chaque paire d'échantillons à travers leurs listes de gènes respectives. Nous avons utilisé un test statistique d'eMADGene permettant d'étudier la composition en gènes de deux régions r_1 et r_2 issus de deux échantillons distincts e_1 et e_2. Soit $n_{1,2}$ le nombre de gènes communs entre r_1 et r_2. Nous considérons comme hypothèse nulle une indépendance des échantillons. Par conséquent l'ordre des gènes d'un échantillon est indépendant de celui d'un autre échantillon. A partir de cette hypothèse, il est possible d'estimer la probabilité que la région r_2 comporte au moins $n_{1,2}$ gènes communs avec ceux de r_1, compte tenu du nombre de gènes communs entre e_1 et e_2 et des tailles des ensembles étudiés.
Dans notre exemple, les régions analysées sont systématiquement la première moitié du classement (gènes sous-exprimés : les résultats sont équivalents si l'on considère la deuxième moitié du classement). La Fig. 29 montre la matrice carrée des p-valeurs obtenues. On observe une très nette différentiation des classes d'échantillons. Ce résultat indique que la pathologie affecte considérablement le transcriptome et que cet effet est suffisamment global pour être identifiable sans sélection préalable de gènes discriminants.

La figure indique par ailleurs un autre fait remarquable : les résultats obtenus sont d'une qualité comparable à travers tous les échantillons quelles que soient leurs provenances (intra-études ou inter-études).

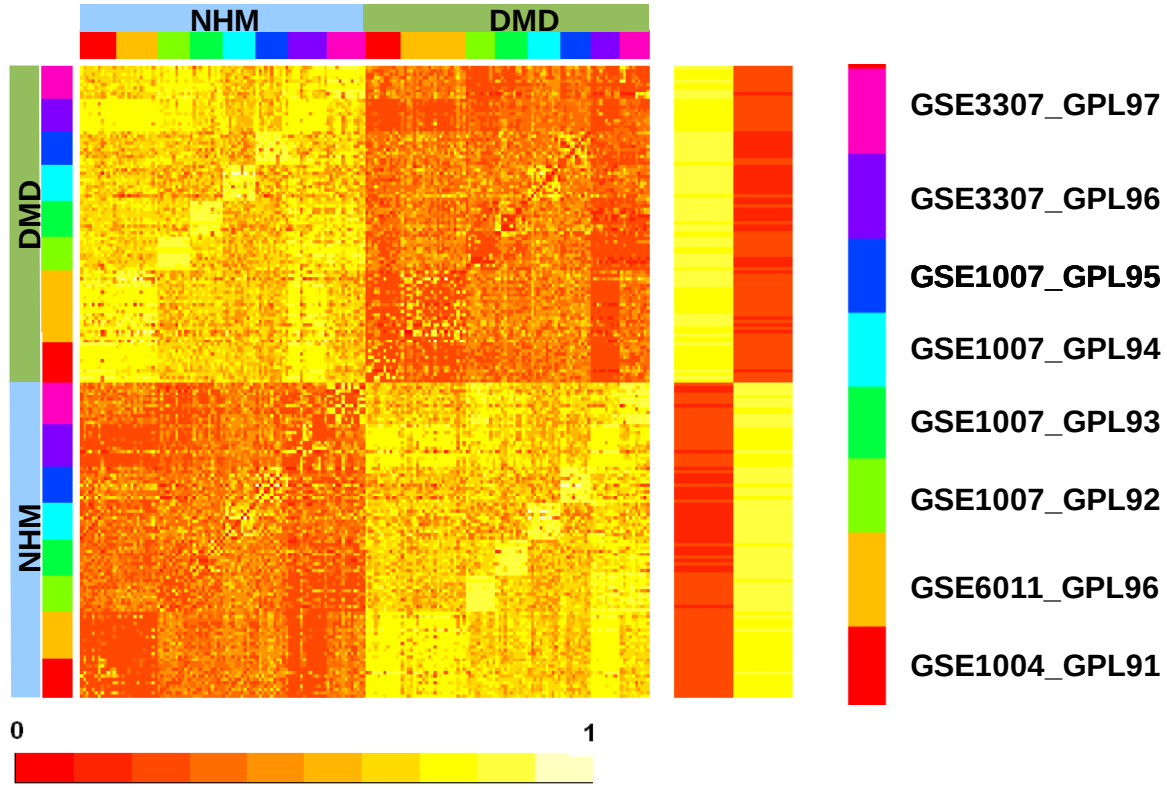


Fig. 29. *Méta-analyse de données publiques sur la maladie de Duchenne.*
La matrice de gauche montre les intensités de p-valeurs selon un code couleur indiqué en bas de la figure, allant de 0 à 1 en dix niveaux équirépartis. Les deux colonnes centrales correspondent aux p-valeurs calculées à partir de listes de gènes discriminants (à gauche, les gènes surexprimés dans les échantillons DMD ; à droite les gènes surexprimés dans les échantillons normaux). La légende de droite indique la provenance des échantillons (études et plateformes de GEO).

Nous avons utilisé un autre test statistique d'eMADGene pour extraire les gènes plus fréquemment surexprimés (deuxième moitié du classement) dans une classe d'échantillon que dans une autre sur la totalité des 193 échantillons. Nous avons appliqué une méthode permettant de rechercher les gènes ayant la plus faible p-valeur dans un test exact de Fisher (1000 gènes ont été retenus pour notre analyse). Nous obtenons ainsi deux listes : une liste de gènes dont la surexpression est remarquablement plus fréquente dans les échantillons DMD et une liste de gènes surexprimés préférentiellement dans les échantillons sains (NHM). Nous avons croisé ces deux listes avec les échantillons à notre disposition (centre de la Fig. 29), en appliquant le même type de test que celui utilisé pour la matrice carrée. Dans ce test, on compare la deuxième moitié du classement avec les deux listes de références. On remarque que les échantillons DMD (resp. NHM) ont une forte proportion de gènes appartenant à la liste DMD (resp. NHM) dans la partie surexprimée. Ce résultat en soi n'est guère surprenant étant donné la méthode de création de ces listes. On observe cependant que cette sélection de gènes améliore notablement la discrimination : les p-valeurs obtenues permettent de distinguer sans ambiguïté les deux classes considérées. En comparaison avec la matrice carrée, on s'aperçoit en effet que le bruit induit par des gènes non discriminants est considérablement réduit.

6.4. Application à la comparaison d'études

Cette deuxième application repose sur la recherche de clusters communs à deux études indépendantes. Chaque étude comporte un ensemble d'échantillons et est associé à un ensemble de gènes spécifiques. La classification hiérarchique ascendante permet de regrouper les gènes selon leur similarité basée sur une mesure de corrélation. Elle donne lieu à une liste ordonnée des gènes qui peut être analysée par eMADGene. A partir de deux listes issues de la classification, le logiciel recherche des régions comportant des biais de représentation : ces régions comportent un nombre surprenant de gènes communs qui sont de fait corrélés dans les deux études indépendantes. Cette méthode permet ainsi de découvrir des clusters communs entre études.

Une alternative à la classification hiérarchique consisterait à utiliser des logiciels permettant d'optimiser l'ordre des gènes de façon à maximaliser la similarité des gènes adjacents. Certaines méthodes ont été notamment proposées pour réorganiser l'ordre des gènes de sorte à mieux regrouper les gènes similaires ([Bar-Joseph et. al. 2001], [Caraux et. al. 2005]).

Pour mettre en évidence la capacité intégrative de notre outil, nous avons comparé deux études portant sur deux espèces distinctes, l'homme et la souris :

- L'étude sur l'homme (GSE 1004, GPL91) fait partie des études utilisées dans la section précédente et comporte des échantillons de muscles sains et DMD.
- L'étude sur la souris (GSE 7187, GPL 339) a pour objet d'examiner le profil d'expression des gènes sur des souris saines et des souris ayant subi une mutation entraînant une pathologie musculaire appelée mdx. Cette maladie s'apparente à la dystrophie DMD chez l'homme. Une troisième catégorie d'échantillons porte sur des souris transgéniques exprimant de l'utrophine (le gène correspondant est une des cibles thérapeutiques de la maladie de Duchenne).

Nous avons procédé à un filtrage des données d'expression, une normalisation LOWESS ainsi qu'à un centrage médian des profils d'expression. Les deux études ont été ensuite classées selon la méthode hiérarchique ascendante développée dans le logiciel Cluster (mesure de similarité : corrélation non centrée ; regroupement par centroïdes). Le résultat d'une classification hiérarchique ascendante est un dendrogramme dans lequel les gènes sont ordonnés de sorte à favoriser le rapprochement des gènes les plus corrélés (l'heuristique d'ordonnancement utilisée par le logiciel Cluster est décrite dans [Eisen 1998]). Une signature commune entre deux études (ensemble de gènes co-exprimés) devrait donc apparaître sous la forme d'une région commune (par région, on entend une sous-liste contiguë d'une liste ordonnée).

Une méthode d'eMADGene permet de rechercher dans deux listes d'identifiants des régions similaires. La correspondance entre régions est basée sur la sur-représentation de certains gènes : leur présence dans chacune des régions est remarquablement plus importante que celle attendue compte tenu de la taille respective des deux listes ainsi que de la taille des régions.

L'utilisateur doit spécifier en entrée la taille d'une fenêtre de recherche : cette taille définit celle des deux régions analysées. La fenêtre de comparaison balaie l'ensemble des positions possibles sur les deux listes (un paramètre permet de définir le pas d'analyse (pas de décalage de la fenêtre) pour accélérer le traitement dans le cas de listes de très grandes tailles).

L'algorithme de recherche procède dans une première étape par la recherche de la correspondance optimale entre les deux listes. Cette dernière est quantifiée par une p-valeur. Si cette estimation est inférieure à un seuil fixé par l'utilisateur, alors les deux régions correspondantes sont étiquetées par un numéro (indice de cluster). La deuxième étape de l'étape consiste à élargir la région de

correspondance en faisant glisser de manière symétrique les deux fenêtres à partir de la position initiale de part et d'autre de cette position. Tant que les p-valeurs de ces nouvelles positions sont inférieures au seuil, les positions sont étiquetées en utilisant le même indice de cluster. Cette étape terminée, l'algorithme réitère les deux étapes tant qu'il trouve des positions non encore étiquetées et possédant une p-valeur suffisante, ou tant qu'un nombre maximal de clusters fixé par l'utilisateur n'a pas été atteint.

En sortie de cette méthode, l'utilisateur obtient deux listes de tailles identiques à celle des deux listes de gènes. A chaque position est associé un indice de cluster. La Fig. 30 montre le résultat obtenu sur les deux études précédemment décrites. La meilleure correspondance entre régions est indiquée par les flèches en rouge. Cet optimum concerne aussi bien la taille du cluster que la p-valeur associée (de l'ordre de 10^{-34} contre 10^{-11} pour le cluster suivant). Fait remarquable, cette région correspond précisément à des gènes discriminants, comme le montre la signature des deux régions (les échantillons ont été classés dans le même ordre : classe saine en premier, et malade ensuite). Cette signature correspond à des gènes surexprimés dans le tissu malade (valeur d'expression apparaissant en rouge) par rapport au tissu sain (valeur d'expression apparaissant en vert).

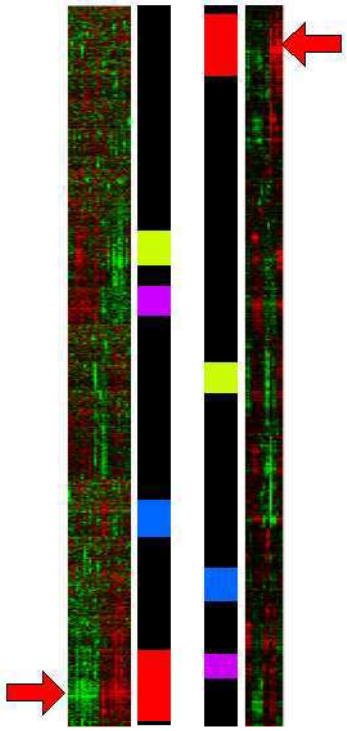


Fig. 30. *Extraction de régions similaires par comparaison de deux études.*
Les deux colonnes extérieures montrent les deux classifications hiérarchiques ascendantes. Celle de gauche est relative à l'étude humaine GSE 1004 et celle de droite au modèle murin (GSE 7187). La taille de la fenêtre d'analyse a été fixée à 200 et le pas de recherche à 20. Le seuil de p-valeur maximale a été fixé à 10-5 et le nombre maximal de clusters à 5.

6.5. Conclusion

Cette étude a montré l'intérêt de la base de données MADGene. Son interface web est conçu comme un outil d'intégration pour les utilisateurs. Il permet de convertir les identifiants en offrant

des passerelles entre un référentiel et un autre. Cet outil permet de croiser des expérimentations entre espèces différentes comme l'a démontrée l'application précédente. Le logiciel eMADGene est une application qui permet de faciliter ce travail d'intégration, selon une approche consistant à s'affranchir des valeurs numériques pour effectuer des traitements basés uniquement sur des listes ordonnées d'identifiants. Cette méthode possède l'avantage de contourner les nombreux obstacles liés à l'hétérogénéité des expériences et des technologies. Les exemples illustratifs ont permis de montrer la pertinence d'une telle approche, en révélant des phénomènes robustes et validées par des études indépendantes.

L'apport majeur de la base de données MADGene est de proposer à la communauté des chercheurs un outil de conversion dont la couverture est supérieure à celle des autres outils disponibles. Le nombre de types d'identifiants acceptés est supérieur, ainsi que le nombre d'espèces traités. En outre, cet outil est convivial et simple d'emploi. Ma contribution principale dans ce projet a été de concevoir l'application eMADGene. Ce logiciel permet de pratiquer en toute facilité des études impliquant de multiples plateformes technologiques et portant sur des espèces diverses. Les opérateurs statistiques fournis permettent à l'utilisateur d'explorer des propriétés communes entre des jeux de données hétérogènes.

Chapitre 7. Vers la méta-analyse de données de puces à ADN

En biologie, la méta-analyse consiste à étudier conjointement plusieurs études indépendantes relatives à un même problème [O'Rourke 2007]. Elle est très utilisée sur des données cliniques établies par des équipes médicales différentes, dans le but d'accroître le nombre d'échantillons et d'améliorer ainsi la puissance statistique de l'étude. Ce chapitre présente une méta-analyse de données d'expression à très grande échelle, intégrant de nombreuses études. Tout comme le chapitre précédent, ce travail associe une démarche scientifique à des développements technologiques et scientifiques, offrant à la fois un service à la communauté des chercheurs ainsi que des méthodes originales d'analyse bio-informatique.

Malgré le volume croissant d'expériences portant sur le transcriptome et de nombreuses publications dans ce domaine, l'exploitation de ces données reste très malaisée. Ce phénomène s'explique par le manque d'outils et de méthodes pour intégrer cette masse d'informations. J'ai collaboré avec l'équipe INSERM nantaise U915 pour le développement de MADMuscle, une ressource accessible par internet dédiée aux pathologies musculaires. MADMuscle est un outil comportant plusieurs facettes :

- Une collection de données de puces à ADN, extraite à partir de *Gene Expression Omnibus*. Les données ont été traitées selon le protocole décrit dans le chapitre 4.2. L'utilisateur de MADMuscle peut consulter les jeux de données, récupérer les données brutes ainsi que les données traitées. Il peut également visualiser les classifications hiérarchiques ascendantes de ces données et récupérer les fichiers de classification correspondants.
- Une base de données de co-expression des gènes, obtenue à partir de l'extraction de clusters sur les jeux de données. L'utilisateur peut consulter la liste des gènes co-exprimés dans une étude, ainsi que l'annotation fonctionnelle correspondante.
- Un outil d'interrogation des données de MADMuscle à partir d'une liste de gènes défini par l'utilisateur. Cet outil permet de confronter une liste de gènes d'intérêt avec les listes de clusters de MADMuscle. La requête fournit comme résultat des groupes de gènes co-exprimés ayant une intersection statistiquement significative avec la liste de gènes donnée par l'utilisateur. Le chercheur peut ainsi déterminer quelles sont les études révélant des mécanismes de co-expression sur ces gènes d'intérêt. La nature de ces études (type de pathologie étudiée, phénotype des patients,…) peut amener le biologiste à mieux comprendre le fonctionnement de ces gènes.

A partir de *Gene Expression Omnibus*, 533 jeux d'expression ont été collectés ayant trait aux études sur le muscle. Après des étapes de normalisation et de clusterisation, 1 247 clusters de gènes ont été extraits, provenant de 7 organismes et de 116 plateformes technologiques. Ces clusters ont été

annotés fonctionnellement en utilisant *Gene Ontology* et sont disponibles aux chercheurs sur le site : www.madtools.org. L'utilisateur peut rechercher des clusters les plus proches d'une liste de gènes en sa possession, indépendamment de la notion d'espèce ou de plateforme technologique. La comparaison entre la liste utilisateur et la base de clusters se fait par un critère statistique de similarité (test exact de Fisher).

La mise à disposition d'une large collection de données portant sur une pathologie cible ouvre une nouvelle dimension dans l'analyse du transcriptome. Elle permet l'étude d'un large ensemble de gènes exprimés au sein d'une grande diversité de conditions expérimentales. Ce type de méta-analyse est rendue possible par l'amélioration de la qualité des données, de leur précision et de leur reproductibilité [Irizarry 2005]. La méta-analyse de puces à ADN de référence est probablement celle portant sur le cancer de la prostate [Rhodes 2002]. Depuis, de nombreuses études ont été menées, notamment en cancérologie ([Choi 2004], [Wang 2004]).

Peu d'outils sont cependant disponibles pour faciliter le travail d'intégration du biologiste. Deux types d'outils sont apparus, le premier porte sur la comparaison de listes de gènes différentiellement exprimés, le second sur les co-expressions des gènes sur les jeux expérimentaux.

Certains systèmes proposent des services analogues à ceux introduits par MADMuscle. On peut notamment citer deux outils basés sur les listes permettent de récupérer des listes de gènes discriminants : LOLA [Cahan 2005] et L2L [Newman 2005]. Ces outils permettent de croiser une liste utilisateur avec les listes annotées. Un inconvénient de ce type d'approche provient du fait que les listes n'ont pas été établies par le même protocole d'analyse, ce qui biaise la comparaison selon le prétraitement, le test statistique et le seuil de confiance utilisés. D'autres auteurs ont donc privilégié la ré-analyse des données brutes. L'un des services le plus connu est *Oncomine* [Rhodes 2004], une collection de référence sur le cancer, relative à 360 expériences, 25 447 échantillons et 40 types tumoraux. L'outil permet à l'utilisateur de rechercher des gènes différentiellement exprimés pour une pathologie donnée ou de comparer des sous-types de cancer. Il est cependant difficile de connaître les gènes partageant des signatures d'expression identiques sur plusieurs expériences sans de fastidieuses interventions manuelles. Un autre outil, *GeneChaser* [Chen 2008], a ré-annoté automatiquement 1 515 jeux de données de GEO pour établir des listes de gènes différentiellement exprimés. CMAP [Lamb 2006] et EXALT [Wu 2009] offrent des services similaires.

Le deuxième type d'outil est basé sur l'analyse de la co-expression. A partir d'un nombre suffisant de conditions expérimentales et pour peu que celles-ci soient de natures différentes, des mécanismes de co-régulations peuvent être détectées grâce aux profils d'expression [Eisen 1998]. La robustesse de telles analyses a été démontrée [Lee 2004]. La méta-analyse de référence basée sur des co-expressions est certainement l'étude portant sur 1 975 jeux de données et 22 types de tumeurs [Segal 2004]. L'extraction de 2 849 ensembles de gènes a permis de mettre en évidence des groupes de gènes évoluant de façon cohérente sur différents jeux et pathologies.

Tous les chercheurs souhaitant pratiquer une méta-analyse doivent entreprendre un effort fastidieux de collecte d'informations et de normalisation des données. La première section de ce chapitre décrit comment a été construite cette collection de données, accessible à tous, dans un objectif de mise à la disposition d'un support cohérent d'un large corpus d'études de transcriptome menées sur les pathologies musculaires. Les deux sections suivantes montrent comment nous avons exploité cette base de données pour nos propres travaux.

7.1. Description de l'outil MADMuscle

La stratégie que nous avons suivie pour la méta-analyse de pathologies musculaires reprend les concepts décrits précédemment, en utilisant dans un premier temps les profils d'expression qui ont l'avantage de révéler des co-régulations de manière générale et non uniquement tributaire de classes connues a priori. Le principe de comparer des listes de gènes a également été retenu. Il permet en effet de rechercher des clusters similaires à une liste de gènes d'intérêt, une fonctionnalité qui fait tout l'intérêt d'un outil comme MADMuscle. La démarche que nous avons observée poursuit notamment les objectifs suivants :

- l'identification de signatures robustes pour une pathologie sur des études indépendantes,
- la découverte de groupes de gènes partageant potentiellement un même mode de fonctionnement sur différents états pathologiques ou liés à des traitements cliniques,
- l'extraction d'ensembles de gènes co-exprimés sur des modèles humains et animaux.

L'outil MADMuscle repose sur une collection de données de puces à ADN. Nous allons décrire la constitution de cette base de données ainsi que les méthodes appliquées pour sa création : la normalisation des jeux de données publiques, la clusterisation hiérarchique de ces jeux et enfin l'extraction de clusters.

La base de données MADMuscle

Elle regroupe toutes les données transcriptomiques relatives à des études portant sur les pathologies musculaires extraits de GEO. Ces données ont été identifiées à partir des mots-clés suivants : "muscle", "myo", "heart" or "cardio". La base représente un total de 535 jeux analysés correspondant à 447 séries uniques de GEO, 116 différentes plateformes de puces et 7 différentes espèces (*Homo sapiens Hs, Mus musculus Mm, Rattus norgegicus Rn, Canis familiaris Cf, Gallus gallus Gg, Drosophila melanogaster Dm, and Caenorhabditis elegans Ce*).

Normalisation des données

La méthode des k plus proches voisins a été appliquée dans le cadre de la normalisation des données de puces à ADN pour évaluer les valeurs manquantes dans les jeux de données [Troyanskaya 2001]. Les effets non-linéaires comme le bruit de fond ou la saturation ont été corrigés par LOWESS [Yang 2002], utilisant une procédure de normalisation canal par canal [Workman 2002], chaque échantillon étant normalisé par rapport au profil médian de tous les échantillons.

Clusterisation hiérarchique

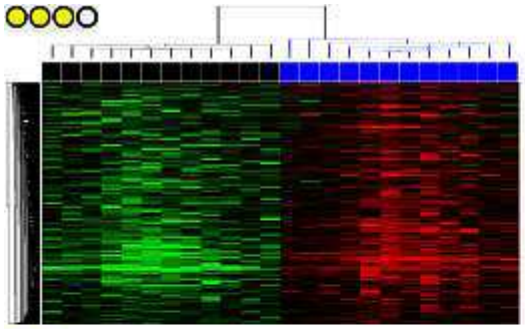
Une classification hiérarchique ascendante a été appliquée [Eisen 1998] pour permettre aux utilisateurs d'analyser les co-expressions sur les données brutes ainsi que sur les données normalisées. Une clusterisation par saut moyen (*Average Linkage*), introduisant la mesure de similarité par corrélation de Pearson, a été effectuée au moyen du logiciel Cluster 3.0 [Hoon 2004] Les profils d'expression ont subi une transformation logarithmique et ont été centrés sur la médiane du profil.

Extraction des clusters

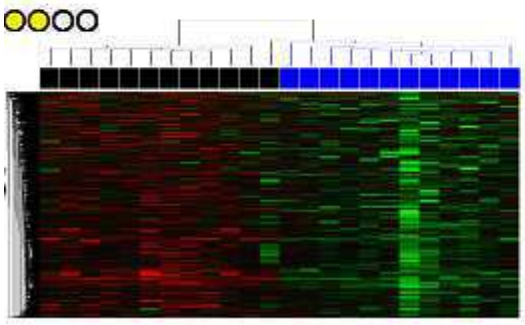
Les jeux d'expression contenant moins d'une dizaine d'échantillons ont été rejetés. L'extraction des clusters est donnée par un algorithme basé sur la stabilité des partitions issues de multiples itérations de k-means (méthode dite des k-means stables [Tavazoie 1999]). Un millier d'exécutions indépendantes de k-means a été lancé, en utilisant des initialisations aléatoires des centroïdes (le nombre maximal de clusters a été fixé à k=9). Une matrice m de co-occurrences a été définie pour chaque paire de gènes. L'élément $m_{i,j}$ correspond à la fréquence de la collocation des gènes g_i and g_j dans un même cluster. Un graphe G de co-occurrences a été établi à partir de m. Les sommets de G correspondent aux gènes ; les arcs correspondent aux paires de gènes dont la fréquence $m_{i,j}$ est supérieure à un seuil fixé à 0.95. L'algorithme d'extraction des clusters consiste à sélectionner les composantes connexes de G.

Pour chaque cluster, un indice de qualité a été déterminé. Cet indice est basé sur une mesure statistique de la pertinence des corrélations au sein d'un cluster. Pour chaque paire de gènes, connaissant une valeur de coefficient de Pearson et le nombre d'échantillons pris en compte, il est possible de définir une p-valeur basée sur une statistique de Student. En considérant toutes les paires de gènes d'un cluster, on peut définir la qualité du cluster par une moyenne géométrique des p-valeurs. Pour les clusters de grande taille, une estimation de cette mesure est effectuée par une technique de ré-échantillonnage. Par extension, nous avons défini la qualité d'un jeu d'expression par la moyenne arithmétique de la qualité de ses clusters. MADMuscle contient 1 247 jeux de bonne ou très bonne qualité sur les 4 400 jeux au total.

La Fig. 31 présente deux exemples de clusters extraits à partir d'un jeu de données (jeu portant le code GSE1004_GPL91 dans la base de données GEO). Ce correspond à une étude portant sur la maladie de Duchenne [Haslett 2002]. La Fig. 31 indique deux clusters de MADMuscle issus de cette étude, formant des groupes de gènes différentiellement exprimés entre tissus sains et tissus DMD. Le cluster du haut (surexpression du muscle DMD) contient des gènes que l'on identifie par le symbole DMD+. Le second cluster caractérisé par une sous-expression du muscle DMD sera noté DMD-.



GO ID	GO term	Number in array	Number in cluster	Enrichment	p-value
5581	collagen	26	12	3.744	0.0000
7160	cell-matrix adhesion	39	15	3.12	0.0000
9611	response to wounding	199	45	1.834	0.0000
44420	extracellular matrix part	57	22	3.131	0.0000
44421	extracellular region part	345	68	1.599	0.0000



GO ID	GO term	Number in array	Number in cluster	Enrichment	p-value
6413	translational initiation	23	11	5.037	0.0000
6417	regulation of translation	32	13	4.279	0.0000
6446	regulation of translational initiation	17	9	5.576	0.0000
9889	regulation of biosynthetic process	77	20	2.736	0.0000
44424	intracellular part	2785	324	1.225	0.0000

Fig. 31. *Deux clusters issus d'une étude sur la maladie de Duchenne.*
Les images de gauche représentent une classification hiérarchique des gènes des groupes DMD+ et DMD-. Les patients sont classés en colonne (normal : noir ; DMD : bleu) et les gènes en ligne. Les cercles jaunes indiquent une échelle de qualité des clusters définie à partir d'une discrétisation de la p-valeur moyenne : le cluster du haut (3 : bonne qualité) est de meilleure qualité que celui du bas (note 2 : assez-bonne qualité). Le fait que le groupe DMD+ soit plus compact que le DMD- s'est vérifié par méta-analyse. Les tables de droite montrent l'annotation fonctionnelle issue de ces clusters.

Cet exemple montre l'intérêt que peut avoir un chercheur à explorer MADMuscle pour valider ses propres travaux. Comme indiqué précédemment, un outil d'interrogation permet d'identifier les clusters de MADMuscle similaires à une liste de gènes établi par l'utilisateur. La connaissance de ces clusters permet de renforcer la pertinence d'une étude en la confrontant aux résultats d'autres chercheurs. A titre d'illustration, nous avons effectué deux études : la première est une méta-analyse sur la maladie de Duchenne et la deuxième est l'identification de méta-clusters.

7.2. Méta-analyse portant sur la maladie de Duchenne

Nous allons montrer dans cette section l'intérêt que peut présenter l'outil MADMuscle pour un biologiste. L'application visée concerne l'étude des gènes impliqués dans la maladie de Duchenne. L'utilisateur souhaite confronter une liste de gènes identifiés par sa propre étude aux informations potentielles contenues dans les études déjà publiées dans ce domaine. Cette analyse permet à l'utilisateur de valider ses résultats : la fiabilité des résultats est renforcée par l'intégration de multiples jeux de données.

Nous avons pris comme support d'étude personnelle une étude DMD qui nous a permis d'extraire deux listes de gènes au fort pouvoir discriminant (liste DMD+ de 483 gènes et liste DMD- de 473 gènes). La consultation de l'outil MADMuscle a conduit à l'identification de 5 études portant sur la même pathologie. La méta-analyse porte sur 5 études DMD indépendantes de GEO (GSE465, GSE1004, GSE1007, GSE3307 et GSE6011). Une méta-matrice a été construite à partir des 5 matrices initiales pour définir des méta-profils d'expression. Un test de Student a permis d'extraire les gènes différentiellement exprimés. La comparaison avec les listes de gènes a conduit à ne

valider qu'un sous groupe de gènes (202 DMD+ ; 109 DMD-). Le fort taux de gènes invalidés s'explique par le fait que l'étude initiale comportait un nombre restreint d'échantillons DMD. Pour déterminer la cohérence de ces deux listes, nous avons recherché dans les jeux de MADMuscle le nombre de collocations d'une paire de gènes dans le même cluster. Afin de ne pas biaiser cette mesure, nous avons considéré l'existence d'une mesure effective de ces gènes (les puces ne comportent pas le même ensemble de gènes) et calculer la probabilité d'obtenir un tel niveau de collocations.

La Table 16 montre un exemple simplifié des informations nécessaires pour une analyse des collocations. Il est important de ne prendre en considération que les expériences pertinentes. Dans l'exemple ci-dessous, les gènes g_1 et g_2 n'ont été en fait mesurés conjointement que sur trois études : GSE1, GSE4 et GSE5 (on peut supposer que les autres études reposent sur des puces ne contenant pas les deux gènes). Sur un total de neuf clusters, deux cooccurrences sont observées (GSE1-C1 et GSE4-C4). Compte tenu de ces observations, il est possible de calculer une p-valeur de cooccurrences en considérant que les gènes sont indépendants entre eux.

	Etudes de Gene Expression Omnibus													
	GSE1			GSE2		GSE3			GSE4				GSE5	
	clusters													
	C1	C2	C3	C1	C2	C1	C2	C3	C1	C2	C3	C4	C1	C2
g_1	1	0	0	X	X	0	1	0	0	0	0	1	0	0
g_2	1	0	0	1	0	X	X	X	0	0	0	1	1	0

Table 16. *Occurrences des gènes g_1 et g_2 dans les 5 études sélectionnées. L'étude GSE1 possède 3 clusters C1 à C3; le chiffre 1 indique que g_1 est présent dans le cluster C1 de cette étude. Le chiffre 0 indique une absence et le symbole X désigne l'absence de mesure : le gène g_1 n'est pas mesuré dans l'étude GSE2.*

A partir de cette p-valeur, exprimée en valeur négative du logarithme, on détermine une mesure de dissimilarité. Il est dès lors possible de visualiser la similitude entre gènes en utilisant la technique de *multidimensional scaling* (MDS [Cox 2001]). Cette méthode projette les gènes sur un espace de telle sorte que les distances entre gènes respectent approximativement leurs dissimilarités.

La Fig. 32 représente le résultat du MDS sur les listes DMD+ et DMD-. On observe une bonne séparation entre les deux groupes de gènes. Ce résultat conforte l'intérêt des deux listes de gènes que nous avions constituées. Rappelons que ces listes ont été extraites à partir d'une étude particulière portant sur la maladie de Duchenne. La Fig. 32 a été obtenue à partir d'un très vaste éventail d'études portant sur des pathologies variées et des espèces distinctes. La séparation des deux groupes signifie que les gènes se retrouvent de façon générale dans des clusters de co-expression spécifiques à la classe auxquels ils appartiennent. Ce fait indique que ces groupes de gènes possèdent leur propre cohérence comportementale et suivent des mécanismes de régulation particuliers.

Comme le montre la figure, on observe également une sur-représentation de la fonction *adhésion cellulaire* dans le groupe DMD+. L'adhésion cellulaire joue un rôle capital dans le maintien et le fonctionnement du tissu musculaire. Il en est de même pour les gènes impliqués dans la *région extracellulaire*. Concernant le groupe DMD-, la fonction G*ene Ontology* sur-représentée est la *mitochondrie*. Un disfonctionnement mitochondrial entraîne irrémédiablement un fonctionnement délétère du tissu musculaire.

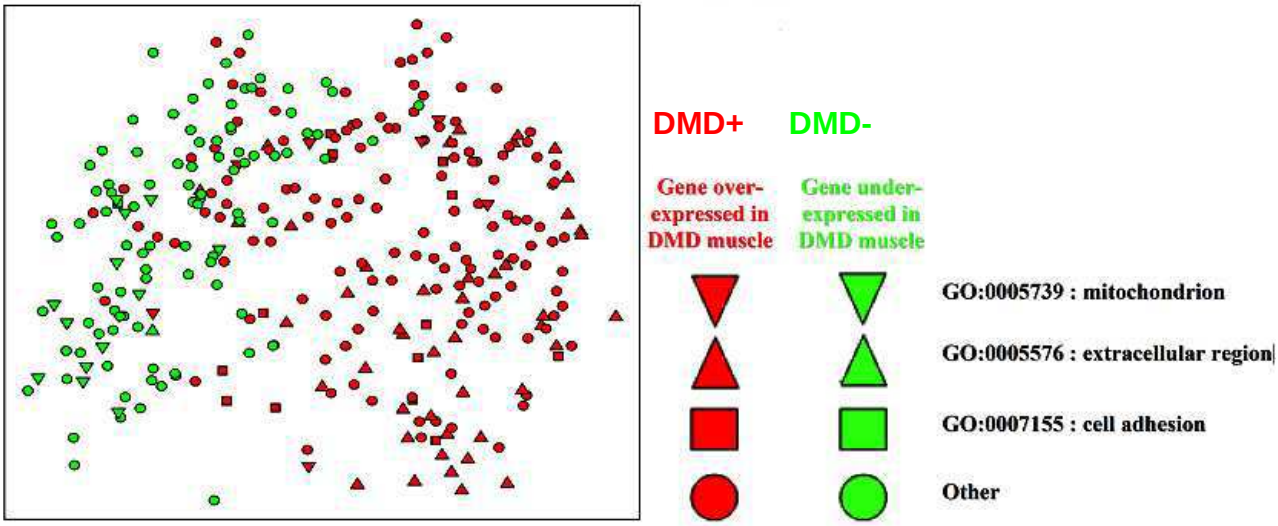


Fig. 32. *Représentation par multidimensional scaling des groupes DMD+ et DMD-.*

7.3. Identification de méta-clusters et annotation fonctionnelle des pathologies musculaires

Cette deuxième illustration de l'intérêt de l'outil MADMucle est une étude globale du comportement des gènes impliqués dans les pathologies musculaires. Contrairement à l'analyse précédente ciblée sur une maladie particulière et des listes de gènes spécifiques, nous recherchons ici à regrouper de manière systématique l'ensemble des signatures (clusters de co-expression) de la collection de MADMucle. L'objectif est de constituer une "carte" générale des signatures afin d'identifier les fonctions activées dans les pathologies musculaires (comme par exemple la réponse immunitaire).

Cette deuxième étude est basée sur une application systématique de la mesure de similarité de l'outil d'interrogation de MADMuscle. A partir des N=1 247 clusters de bonne qualité de la base, nous avons constitué une matrice de similarités. Cette matrice $m_{i,j}$ de taille NxN définit la similarité entre un cluster c_i et un cluster c_j de la base. Cette mesure est déterminée par un test exact de Fisher basé sur la proportion relative des gènes de c_i dans c_j par rapport à la proportion observée dans la plateforme associée.

La Fig. 33.A permet de visualiser cette matrice de similarité entre clusters. La classification hiérarchique ascendante (CHA) opérée sur la matrice à l'aide du logiciel Cluster 3.0 [Hoon 2004] a permis un réordonnancement des indices en lignes (la matrice étant symétrique, nous avons opéré une CHA sur les lignes et appliqué ensuite le même ordre sur les colonnes). Nous avons analysé manuellement cette matrice et avons relevé 6 méta-clusters importants (M1 à M6), chacun d'eux contenant au moins 15 clusters provenant de diverses études, plateformes ou espèces.

La Fig. 33.B est une représentation originale des fonctions biologiques pouvant être associées aux clusters. L'axe des *y* est relatif aux 7 554 fonctions de *Gene Ontology* (GO), réordonnés par CHA de façon à rapprocher les fonctions GO les plus similaires. La classification hiérarchique regroupe ainsi les vecteurs de p-valeurs formées par annotation d'une fonction GO pour chacun des clusters. Le relief représente l'intensité des p-valeurs (plus celle-ci est petite, et plus l'altitude (axe *z*) est importante). L'échelle des couleurs est relative à l'enrichissement. Cette notion représente le rapport entre la proportion de gènes ayant la fonction GO dans le cluster relativement à l'ensemble des gènes de la puce. Si ce rapport est supérieur à 1, il y a surreprésentation de la fonction ; s'il est

inférieur à 1, une sous-représentation est observée. Ce dernier cas n'est pas relevé dans cette étude, sauf pour des p-valeurs peu significatives.

La Fig. 33.B indique les principales fonctions biologiques extraites. Elles sont effectivement spécifiques des études sur le muscle et on remarque que les pics apparaissent globalement sur les méta-clusters. Ce fait est particulièrement intéressant car il confirme un fait généralement admis : quand plusieurs clusters partagent une même signature, ce pattern commun est plus à même de relever un processus biologique spécifique [Segal 2004]. Cette propriété renforce l'intérêt de l'intégration de données de puces combinant plusieurs plateformes [Zhou 2005] et des analyses multi-espèces [Stuart 2003].

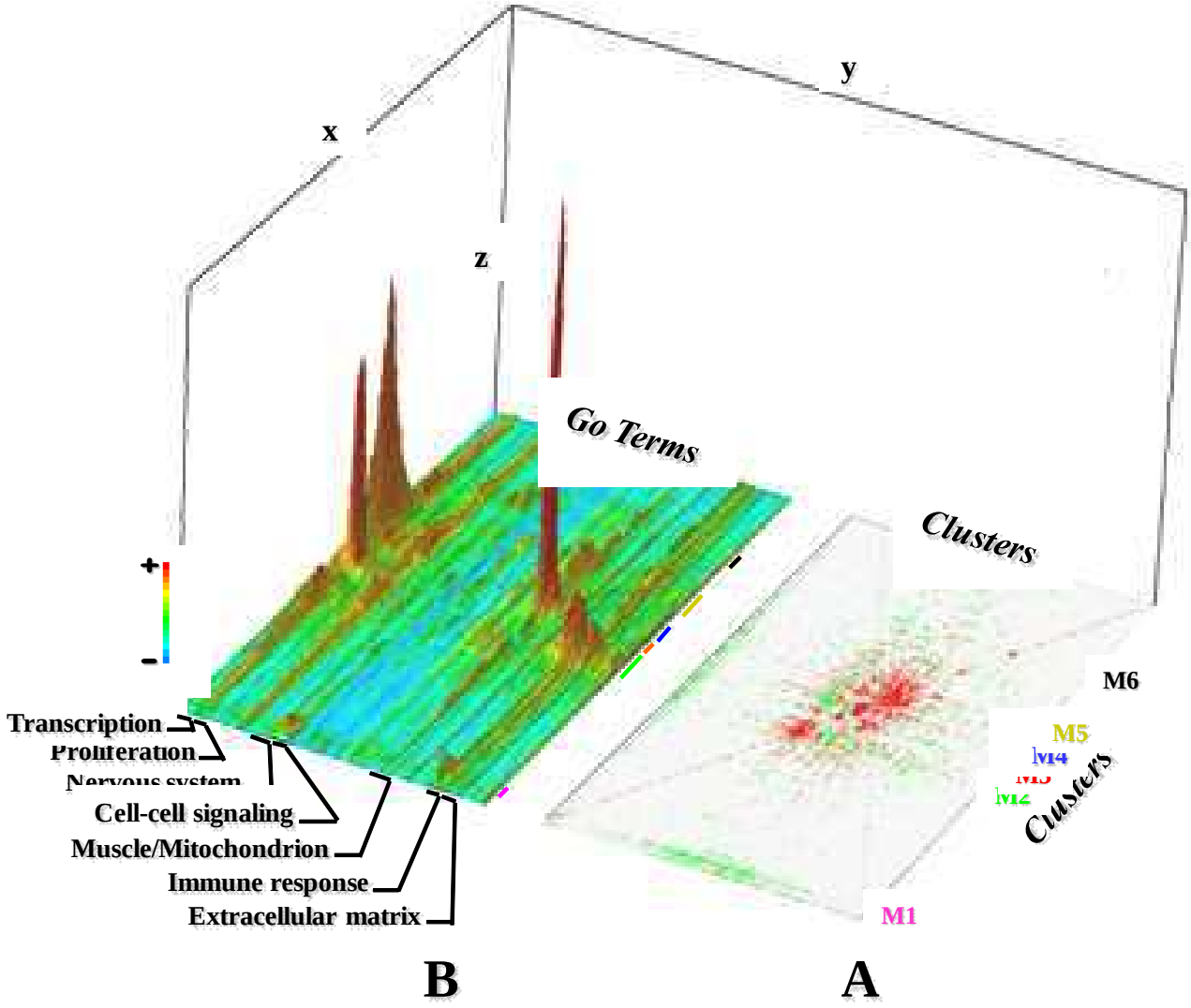


Fig. 33. *Matrice de similarités entre clusters.*

A- La partie droite de la figure représente une matrice symétrique de p-valeurs. Les lignes et colonnes de cette matrice correspondent aux clusters extraits de MADMuscle. La valeur p(i,j) correspond à la probabilité d'observer autant de gènes communs entre le cluster i et le cluster j, en tirant aléatoirement les ensembles i et j parmi l'ensemble des gènes mesurés sur la puce. La matrice est ordonnée au moyen d'une classification hiérarchique pour mettre en évidence des similarités de composition des clusters. Les valeurs en rouges indiquent de fortes similarités, les valeurs en vert indiquent de fortes dissimilarités, les valeurs en gris indiquent de faibles recouvrements entre clusters. Cette représentation a permis d'identifier six groupes de clusters (M1 à M6).

B- La partie gauche de la figure est une représentation 3-D de l'annotation fonctionnelle. L'axe x représente les clusters définis précédemment, l'axe y les annotations définies par Gene Ontology (GO). L'axe z représente la probabilité de rencontrer au hasard autant de gènes du cluster x appartenant à la catégorie fonctionnelle y. Les couleurs représentent l'enrichissement (rapport de la fréquence des gènes appartenant à y dans le cluster sur la fréquence de y dans tout le génome). On observe une remarquable cohésion des classes fonctionnelles et l'émergence de fonctions spécifiques aux méta-clusters M1 à M6.

7.4. Conclusion

L'outil MADmuscle permet de consulter une collection de données relative à des expériences sur les pathologies musculaires. Un regroupement automatique des gènes a été entrepris, qui permet à un utilisateur d'étudier les mécanismes de co-expression dans les jeux de données. Une interface de requête accepte en entrée une liste de gènes d'intérêt et renvoie les clusters dont le contenu en gènes est le plus statistiquement remarquable par rapport à cette liste utilisateur. Cet outil offre ainsi une aide au biologiste désireux de réaliser une méta-analyse, en lui indiquant quels sont les clusters et les expériences similaires à son domaine de recherche. Le biologiste peut très aisément valider ses propres travaux, par comparaison avec des résultats ayant généralement fait l'objet d'une publication.

Mon apport dans cette thématique a essentiellement consisté à utiliser l'outil MADMuscle pour pratiquer des méta-analyses. Plus précisément, j'ai proposé des techniques de visualisation adaptées. J'ai ainsi suggéré une représentation par *multidimensional scaling* pour mettre en évidence des groupes de gènes discriminants à travers l'ensemble des jeux de données. J'ai utilisé pour ce faire une pseudo-distance basée sur un test statistique. J'ai également développé une cartographie en relief pour relever des fonctions biologiques spécifiquement liées aux pathologies musculaires, établissant ainsi des méta-clusters (ensembles de groupes de gènes co-exprimés à travers des jeux très divers, en termes d'espèces, de maladies ou des conditions expérimentales considérées).

Cette étude peut évidemment se généraliser à l'ensemble des données publiques concernant le transcriptome. Elle ouvre la voie à une approche à très grande échelle de ce type de données pour l'étude d'autres pathologies humaines.

Conclusion et perspectives

Mes travaux de recherche au sein de l'équipe COD du Laboratoire d'Informatique de Nantes-Atlantique ont porté sur l'application de techniques de fouille de données à la bio-informatique. Les collaborations menées avec mes collègues biologistes m'ont amené à traiter deux thématiques fondamentales, la classification de protéines et l'analyse de données de puces à ADN.

Le premier axe de recherche a permis de concevoir et de développer des méthodes d'identification d'homologues distants dans des familles de protéines. Ces techniques ont été implémentées par des Séparateurs à Vastes Marges en adaptant la procédure de vectorisation aux séquences protéiques. Une technique originale a été proposée, basée sur une organisation hiérarchique de motifs incluant des classes physico-chimiques d'acides aminés. Plusieurs classifieurs SVM ont été intégrés dans un système de classification reposant sur une agrégation des résultats de ces classifieurs individuels. Les probabilités d'appartenance à la classe recherchée ont été prises en compte comme données d'entrée et l'étude des performances a été axée sur l'aire sous la courbe ROC. La superfamille des cytokines sur laquelle a été appliquée cette méthode est particulièrement importante pour la recherche biomédicale. La collection de protéines qu'elle renferme est très hétérogène. Les tests ont montré la pertinence de cette approche pour la classification d'homologues distants.

Le deuxième axe de recherche porte sur l'analyse des données du transcriptome. J'ai appliqué les concepts et techniques de l'Analyse Statistique Implicative aux données de puces à ADN. J'ai proposé une méthode originale basée sur l'analyse de rangs et utilisé l'Analyse Statistique Implicative pour la sélection de gènes informatifs et pour l'extraction de règles de classification. Ces règles m'ont permis de mettre en œuvre un classifieur qui, malgré sa relative simplicité, s'est montré aussi performant que les techniques les plus efficientes dans ce domaine. J'ai également participé à la mise en place de méthodes d'intégration de données (MADtools) qui permettent de développer de nouvelles pistes d'investigation dans l'analyse du transcriptome. En particulier, la conversion des identifiants entre espèces et plateformes technologiques offre la possibilité de comparer des études hétérogènes. J'ai ainsi pu concevoir et développer eMADGENE, un logiciel d'analyse de listes de gènes. Deux applications ont été réalisées avec cette application : une analyse d'une pathologie musculaire impliquant plusieurs jeux de données et une comparaison de deux classifications hiérarchiques opérées sur un jeu relatif à l'homme et un jeu relatif à la souris. Ces deux illustrations montrent une signature liée à la pathologie musculaire, dont la robustesse est garantie par l'utilisation de jeux indépendants et hétérogènes. Un effort plus systématique de méta-analyse a été entrepris sur les pathologies musculaires disponibles publiquement. J'ai participé à la création de la base de données MADmuscle, accessible comme eMADGENE sur le site des MADtools. Cette base contient un ensemble de clusters de gènes extraits automatiquement à partir des jeux de données. Une interface de requête permet à l'utilisateur de rentrer une liste de gènes d'intérêt. Un outil de MADmuscle recherche les clusters dont le contenu en gènes est similaire, sur la base d'une méthode statistique. Ce type de requête peut grandement faciliter les méta-analyses en déterminant quelles sont les résultats connus dans un domaine de recherche spécialisé.

Concernant le premier axe de recherche, les performances obtenues en classification de protéines sont satisfaisantes ; elles pourraient être certainement encore améliorées, mais il est fort peu probable qu'on puisse atteindre des scores significativement supérieurs. Le problème peut être considéré comme étant en grande partie résolu. Par ailleurs, la thématique de l'identification d'homologues distants suscite peu de travaux à l'heure actuelle, si on se réfère au nombre désormais restreint de publications sur ce domaine. Ces raisons m'incitent à axer mes efforts sur la thématique d'analyse de données du transcriptome. Je vais donc présenter les perspectives à court et moyen termes dans ce contexte de recherche, essentiellement liées à la méta-analyse de données d'expression. Je terminerai par une présentation de certains enjeux cruciaux de la bio-informatique associés à mes perspectives de recherche à long terme. J'y développe notamment une nouvelle manière d'envisager la bio-informatique, à savoir la génomique intégrative. Enfin, il est important de prendre en considération l'évolution des techniques qui bouleversent les habitudes des chercheurs et amènent de nouvelles problématiques. Dans ce cadre, il est difficile d'ignorer les méthodes de séquençage à très haut débit et les défis qu'elles soulèvent pour les bio-informaticiens.

Méta-analyse de données de puces à ADN

Plusieurs axes de recherche peuvent être envisagés pour la méta-analyse. Le premier est une généralisation du travail entrepris sur les pathologies musculaires à l'ensemble des données du transcriptome (du moins des données disponibles publiquement sur Gene Expression Omnibus). Ce travail porte sur la découverte automatique de méta-clusters ainsi que sur leur annotation fonctionnelle. Deux difficultés sont clairement identifiables : la première repose sur la grande taille du nombre de données, équivalent à plusieurs centaines de milliers de clusters, ce qui implique des problèmes d'explosion combinatoire. La deuxième difficulté est relative à la visualisation des informations extraites à l'échelle du génome, ce qui pose des problèmes de représentation et d'aide à l'interprétation pour le biologiste.

Le deuxième axe de méta-analyse concerne l'exploitation de l'outil MADCOW de la plate-forme MADtools. Cet outil repose sur une base de données de corrélation issue de l'ensemble des données GEO. Cette base renferme les paires de gènes corrélés sur les jeux d'expression publiques extraits d'études diverses, tant par l'espèce que par la technologie considérées. Pour rendre les informations comparables, plusieurs seuils de p-valeurs (à partir de 10^{-4}) ont été appliqués, ces derniers étant calculés empiriquement à partir des données. Le but de ce projet est de concevoir et de développer des algorithmes de fouille de données basés sur les graphes de corrélation. La base MADCOW permettrait à l'utilisateur de combiner et de comparer des graphes à partir d'opérateurs dédiés. A titre d'exemple, il serait ainsi possible de visualiser les différences existantes entre deux sous-ensembles de jeux expérimentaux (*e.g.* traitant de 2 types distincts de cancer). On note que cet outil de fouille de graphes peut s'appliquer à d'autres types de graphes que ceux donnés par MADCOW. On peut notamment croiser des graphes de corrélations basés sur les profils d'expression des gènes avec des graphes d'interaction protéiques, en associant un nœud-gène à son équivalent nœud-protéine.

Le troisième axe de recherche porte sur la fouille de données sur les échantillons. Peu d'auteurs ont développé des algorithmes de fouille de données à l'échelle de la puce (le transcriptome mesuré sur cette puce) ; la plupart des travaux reposent sur une approche orthogonale, celle du gène. Considérer la puce comme entité de base permettrait de nouvelles approches. Pour citer un intérêt de ce concept, il serait utile de disposer d'un outil capable de rechercher parmi les jeux de données publiques quels sont les échantillons les plus proches d'un jeu fixé par l'utilisateur. L'apport d'un tel outil pour les biologistes est multiple. Il peut ainsi faire apparaître des similitudes entre profils de patients jusqu'alors inconnues, liant certaines pathologies, et qui révèlent des comportements communs d'expression des gènes. Des raisons plus pratiques de validation sont également envisageables, pour vérifier notamment la cohérence des données d'une expérimentation. Plusieurs difficultés apparaissent dans la mise au point de cet outil, déjà mentionnées par ailleurs, dont notamment le nombre considérable de données (de l'ordre d'un demi-million d'échantillons sur

GEO), posant des problèmes de passage à l'échelle, tant pour l'analyse des données que pour leur représentation.

Génomique intégrative

Une perspective plus large nous amène à envisager le développement de la fouille de données selon un point de vue intégratif. La génomique intégrative est une discipline qui vise à croiser les résultats de plusieurs technologies pour faire de nouvelles découvertes biologiques et pour accroître la fiabilité des résultats.

Le croisement d'informations multiples est certainement le défi majeur en fouille de données génomiques. Les techniques biologiques elles-mêmes se complexifient ; il n'est pas rare de pratiquer une mesure du transcriptome pour étudier les phénomènes suivants :

- Le polymorphisme nucléotidique simple ou *Single nucleotide Polymorphism* (*SNP*). Les SNPs représentent des variations génétiques correspondant à une seule paire de base du génome propre à une espèce. La variabilité du génome des individus est désignée sous le terme de polymorphisme. Ces variations sont fréquentes au sein de la population humaine (de l'ordre d'une paire de base sur mille).
- La variabilité du nombre de copies d'un gène ou *Copy Number Variation* (*CNV*). Cette notion se réfère à une forme particulière de polymorphisme. Ce phénomène correspond à des duplications de gènes au sein du génome. Le nombre de copies peut varier selon les individus d'une même espèce, probablement en raison d'une pression de sélection liée à des causes environnementales.
- La méthylation. La structure de la chromatine (forme de l'ADN dans le noyau) influe sur l'expression génique, l'ADN devant être décompacté pour s'exprimer. La méthylation est un processus épigénétique induisant une modification de l'ADN.
- Les micro-ARNs. Ce sont des régulateurs post-transcriptionnels ayant généralement un effet de silençage des gènes. Ce terme désigne un blocage du gène qui serait exprimé dans des conditions normales.

Il est à noter que certaines approches de génomique intégrative ont été déjà proposées. Le couplage du transcriptome et du génotypage SNP a donné lieu à la technique de l'*eQTL* (*expression quantitative trait loci*); le transcriptome couplé au *ChIP-chip/ChIP-seq* a conduit à l'élaboration de réseaux de régulation (la méthode de *Chromatin ImmunoPrecipitation on Chip* [Ren 2000], trop complexe pour être détaillée dans ce mémoire a pour but essentiel la localisation des sites de fixation de facteurs de transcription).

En termes de fouille de données, la prise en compte de sources hétérogènes permet de répondre à des questions complexes. En ce qui me concerne, la perspective la plus immédiate au vu des travaux déjà menés serait sans aucun doute l'étude de l'activation des gènes par leurs facteurs de transcription. Pour découvrir une telle relation, il faut croiser la connaissance d'une corrélation positive entre un gène et un facteur de transcription (obtenue par méta-analyse de données de puces à ADN) avec celle portant sur la régulation de la transcription. Cette information est donnée par une expérimentation de *ChIP-chip* qui montre que le gène est positif pour ce facteur de transcription. Une analyse plus fine renforcerait cette relation en indiquant qu'un motif spécifique a été découvert sur la région de régulation du gène, cette information nécessitant une analyse de séquence génomique.

Incidences en bio-informatique des nouvelles technologies

Comme nous l'avons mentionné plus haut, le bouleversement des méthodes des chercheurs est souvent lié à l'accroissement des technologies mises en œuvre et la baisse des coûts des expérimentations. De grandes transformations sont en cours, qu'il est utile d'exposer ici.

L'avènement des technologies de *Next Generation Sequencing* (NGS) accélèrera dans la prochaine décennie l'essor de la génomique individuelle et de la médecine personnalisée. Rappelons pour mettre en perspective cet enjeu que le projet *Human Genome* a été initié vingt ans plus tôt. Son objectif était alors le premier séquençage complet du génome humain. Le progrès scientifique a dépassé ce but initial, puisque ce projet a induit un immense effort d'investissement en matière de recherche technologique. Il est fort probable que les technologies NGS vont également bouleverser considérablement le paysage scientifique et révolutionner la génomique. Après l'extraction du génome humain, un projet encore plus ambitieux est en effet en cours d'élaboration. Ce projet intitulé 1000 génomes (*http://www.1000genomes.org/*) consiste à créer une carte des variations génétiques du génome humain. De nombreuses avancées sont envisagées, qui vont permettre par exemple de mieux appréhender les gènes impliqués dans le cancer. Un autre projet phare est sans conteste le projet *Human Microbiome* consistant à séquencer le méta-génome des microbiotes chez l'homme (le microbiome désigne l'écosystème des micro-organismes vivant au sein d'un hôte). Ce projet doit permettre l'analyse du rôle de ces organismes dans la santé et les maladies humaines. Par ailleurs, un projet européen *MetaHIT* a pour but d'étudier le génome de l'ensemble des bactéries constituant la flore intestinale humaine. Les buts affichés sont identiques au projet sur le microbiome : la connaissance des fonctions des bactéries et leurs implications sur la santé.

Le coût décroissant du séquençage des technologies NGS va ouvrir de nouvelles opportunités pour le développement d'applications biologiques et de logiciels bio-informatiques. La masse de données générées par ces nouvelles technologies est impressionnante : de l'ordre du giga-octet par analyse. La gestion et le traitement de ces données sont des véritables défis pour la communauté scientifique.

Cette évolution technologique aura des retombées importantes en analyse bio-informatique. Du succès de celle-ci dépendra une révolution du traitement de la santé humaine. Il sera désormais possible de développer une médecine à la carte (le séquençage du génome d'un patient coûte actuellement environ un millier d'euros avec ces nouvelles technologies).

J'espère avoir montré à quel point le contexte actuel de la génomique fonctionnelle est foisonnant et ouvre d'intéressantes perspectives pour les chercheurs spécialisés dans la fouille de données complexes. Le fait d'avoir pu entretenir une collaboration étroite avec une équipe de biologistes maîtrisant plusieurs techniques d'analyse est un atout qui me paraît incontournable pour mener à bien de telles investigations.

Glossaire

Acide aminé

Composant de base de la protéine.

Acide nucléique

Macromolécule de type *ADN* ou *ARN*.

Codon

Une séquence de trois nucléotides provenant de l'ARN messager et servant de code génétique pour la traduction en acide aminé.

ARN messager (ARNm)

Acide ribonucléique messager. Copie de l'*ADN* utilisée pour la synthèse des protéines.

Chromosome

Les chromosomes contiennent l'information génétique. Ils sont discernables lors de la division cellulaire (23 paires de chromosomes chez l'homme).

Epigénétique

Étude des changements héréditaires dans la fonction des gènes, ayant lieu sans altération de la séquence ADN.

Expression génique

Processus biochimique engendrant la transcription d'un gène.

Gène

Unité d'information sous forme de fragment d'ADN donnant lieu à une synthèse d'ARN.

Génome

Ensemble du matériel génétique d'un individu codé par son *ADN*.

Génotypage

Etude du génome et des variations génétiques pour une population ou une espèce.

Génotype

Information génétique portée par un individu particulier.

Hybridation

Technique utilisant une propriété des séquences nucléiques permettant leur appariement en paire de brins ADN ou ARN.

Next Generation Sequencing

Nouvelle génération de séquençage : ensemble de technologies assurant un séquençage à très haut débit sans la nécessité de procéder à un clonage, et pour un coût moindre que les techniques traditionnelles.

Nucléotide

Acide désoxyribonucléique intervenant dans la composition de l'ADN. Egalement acide ribonucléique pour l'ARN. Sa base azotée forme l'alphabet décrivant une séquence nucléique.

Oligonucléotide

Court fragment d'ADN ou d'ARN obtenu généralement par synthèse chimique.

Phénotype

Ensemble des caractéristiques observables pour un individu donné.

Protéine

Cette macromolécule est la brique de base des organismes vivants. Elle est issue d'un processus en deux étapes : la *transcription* et la *traduction*.

Protéome

Ensemble des protéines exprimées dans une cellule.

Puce

Une grille de sondes déposée sur un substrat (lame de verre, de silicone, …) utilisée pour la détection de séquences complémentaires.

Traduction

Traduction de l'ARN messager (ARNm) en protéine.

Trancriptome

Ensemble des *ARN messagers* produits par un organisme.

Transcription

Processus par lequel un gène associé à la protéine est transcrite en *ARN messager*.

Transcrit

Séquence d'*ARN messager* résultant de la *transcription*.

Liste des acronymes

AA

Acide aminé. Molécule à la base de la composition des protéines.

ADN

Acide désoxyribonucléique. Macromolécule constituant le support de l'information génétque.

ADNc

ADN complémentaire. Copie *ADN* d'un *ARNm* obtenue artificiellement et réalisée pour des raisons techniques (stabilité et facilité d'utilisation).

ARNc

ARN complémentaire d'un autre ARN.

ARNm

Acide ribonucléique messager. Copie de l'*ADN* utilisée pour la synthèse des protéines.

DMD

Duchenne muscular dystrophy. La dystrophie musculaire de Duchenne est une maladie de dégénérescence des fibres musculaires.

GE

Grammatical Evolution. Approche évolutionnaire pour la génération automatique de programmes à partir d'une grammaire.

GEO

Gene Expression Omnibus. Base de données publiques relative au données de génomique fonctionnelle.

GO

Gene Ontology. Ontologie sur les genes selon la composition cellulaire, la fonction moléculaire et le processus biologique.

GPL

Terme de la base de données de *Gene Expression Omnibus* désignant une plateforme technologique (typiquement la liste des gènes déposés sur une puce).

GS

Grammatical Swarm. Essaim grammatical. Génération de programmes à partir d'une grammaire. Utilise une optimisation par essaim de particules (PSO).

GSE

Terme de la base de données de *Gene Expression Omnibus* désignant un jeu de données expérimentales déposé dans GEO.

LOWESS

Locally Weighted Scatterplot Smoothing. Méthode de régression locale utilisée pour normaliser les données d'expression.

MDS

Multidimensional scaling. Méthode projetant des individus dans un espace en tenant compte d'une matrice de dissimilarités.

MDX

Dystrophie musculaire chez la souris.

NGS

Next Generation Sequencing. Nouvelle génération de séquençage : ensemble de technologies assurant un séquençage à très haut débit sans la nécessité de procéder à un clonage, et pour un coût moindre que les techniques traditionnelles.

PSO

Particle Swarm Optimization. Optimisation par essaim de particules. Métaheuristique d'optimisation s'inspirant de comportements sociaux (nuées d'oiseaux, bancs de poissons).

SCOP

Structural Classification of Proteins. Base de données sur les protéines, classées selon leurs structures.

SOV

Segment OVerlap. Mesure de similarité entre structures secondaires de protéines.

SVM

Séparateurs à Vastes Marges ou Machine à Vecteurs de Support. Technique de classification supervisée s’appuyant sur la théorie statistique de l’apprentissage.

Bibliographie

Agrawal, R., Imielinski, T. & Swami, A. (1993). Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216.

Alibes, A. et al. (2007). IDconverter and IDClight: Conversion and annotation of gene and protein IDs. BMC Bioinformatics 8, 9.

Alizadeh A. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature; 403: 503-511.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J, (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96(12) pp. 6 745–6 750.

Altschul, S.F. et al. (1990). Basic local alignment search tool. J. Mol. Biol 215, 403-410.

Ashburner, M. et al. (2000).Gene Ontology: tool for the unification of biology. Nat Genet 25, 25-29.

Bar-Joseph, Z., Gifford, D.K. & Jaakkola, T.S. (2001). Fast optimal leaf ordering for hierarchical clustering. Bioinformatics 17, S22-29.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28, 45-48.

Ballester et al. (2005). Gene expression profiling identifies molecular subgroups amongnodal peripheral T-cell lymphomas. Oncogene.

Benson, D.A. et al. (2000). GenBank. Nucleic Acids Res 28, 15-18.

Bertone, P. & Gerstein, M. (2001). Integrative data mining: the new direction in bioinformatics. IEEE Eng Med Biol Mag 20, 33-40.

Bertucci F. et al. (2002). Profils d'expression génique par puces à ADN dans les lymphomes malins : intérêt pronostique. Bulletin du cancer ; 89 : 7.

Bertucci F. et al. (2005). Gene expression profiling identifies molecular subtypes of inflammatory breast cancers. Cancer Res.; 65(6):2170-8.

Bullinger L., Valk P. Gene Expression Profiling in Acute Myeloid Leukemia. Journal of Clinical Oncology, Vol 23, No 26 (September 10), 2005: pp. 6296-6305.

Bussey, K.J. et al. (2003). MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biol. 4, R27-R27.

Brameier, M. & Banzhaf, W. (2001). Evolving Teams of Predictors with Linear Genetic Programming. Genetic Programming and Evolvable Machines 2, 381-407.

Cahan, P. et al. (2005). List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists. Gene 360, 78-82.

Caraux, G. & Pinloche, S. PermutMatrix (2005). A graphical environment to arrange gene expression profiles in optimal linear order. Bioinformatics 21, 1280-1281.

Carmona-Saez, P. et al. (2006). Integrated analysis of gene expression by association rules discovery. BMC Bioinformatics 7, 54.

de Castro, E. et al. ScanProsite (2006). Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34, W362-365.

Chen, D., Z. Liu, X. Ma, et D. Hua, (2005). Selecting genes by test statistics, Journal of Biomedicine and Biotechnology 2, 132-138.

Chen, R. et al. (2008). GeneChaser: Identifying all biological and clinical conditions in which genes of interest are differentially expressed. BMC Bioinformatics 9, 548.

Chen, R. et al. (2007). AILUN: reannotating gene expression data automatically. Nat Meth 4, 879.

Choi, J.K. et al. (2004). Integrative analysis of multiple gene expression profiles applied to liver cancer study. FEBS Lett 565, 93-100.

Cox, T.F., Cox, M.A.A. (2001). *Multidimensional Scaling*, Chapman and Hall.

Dettling, M. et P. Bühlmann (2002). Supervised selection of genes. Genome. Biol. Res. 3(12), Research, 0069.1-0069.15.

Diaz-Uriarte, R., Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7.

Diehn, M. et al. (2003). SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucl. Acids Res. 31, 219-223.

Ding, C.H. & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349-358.

Edgar, R., Domrachev, M. & Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30, 207-210.

Eisen, M.B. et al. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95, 14863-14868.

Field, D. et al. (2009). 'Omics Data Sharing. Science 326, 234-236.

Furey, T.S. et al. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906-914.

Gasch, A., Eisen M. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy kmeans clustering.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander, (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, pp 531-537.

Gras, R., Couturier, R., Guillet, F, Spagnolo F. (2005). Extraction de règles en incertain par la méthode statistique implicative. In 12èmes Rencontres de la Société Francophone de Classification, Montréal, pp. 148-151.

Gras, R., Diday E., Kuntz P., Couturier R. (2001). Variables sur intervalles et variables intervalles en analyse statistique implicative. In Société Francophone de Classification (SFC'01), Univ. AntillesGuyane, PointeàPître, pp. 166-173.

Gras, R. (1996). L'implication Statistique, Grenoble : La Pensée Sauvage.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol 292, 195-202.

Kearns, M. (1988). Thoughts on hypothesis boosting. Manuscrit non publié.

Kennedy, J., Eberhart, R. C. (1995). Particle swarm optimization. Proceedings of the 1995 IEEE International Conference on Neural Networks (Perth, Australia), IEEE Service Center, Piscataway, NJ, IV: 1942-1948.

Haslett, J.N. et al. (2002). Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle. Proc. Natl. Acad. Sci. U.S.A 99, 15000-15005.

Hatami, N. et al. (2007). Combining Multiple Classifiers: Diversify with Boosting and Combining by Stacking" International Journal of Computer Science and Network Security (IJCSNS), 7(1), 127-131.

Hoon, M.J.L. et al. (2004). Open source clustering software. Bioinformatics 20, 1453-1454.

Hua, S. & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol 308, 397-407.

Huang, D. et al. (2008). DAVID gene ID conversion tool. Bioinformation, 2(10), 428-430.

Hubbell, E., Liu, W. & Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics 18, 1585-1592.

Irizarry, R.A. et al. (2005). Multiple-laboratory comparison of microarray platforms. Nat. Methods 2, 345-350.

Jaakkola, T., Diekhans, M. & Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. J. Comput. Biol 7, 95-114.

Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846-856.

Kennedy, P. et al. (2008). Integrative Visual Data Mining of Biomedical Data: Investigating Cases in Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia. Visual Data Mining 367-388.

Kuncheva, L., Whitaker, C., Shipp, C. & Duin, R. (2003). Limits on the majority vote accuracy in classifier fusion. Pattern Analysis & Applications 6, 22-31.

Lamb, J. et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313, 1929-1935.

Lapointe J. et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. PNAS Jan 20; 101(3): 811.

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

Lee, H.K. et al. (2004). Coexpression analysis of human genes across many microarray data sets. Genome Res 14, 1085-1094.

Leslie, C., Eskin, E. & Noble, W.S. (2002). The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput 564-575.

Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S. (2004). Mismatch string kernels for discriminative protein classification. Bioinformatics 20, 467-476.

Leung, Y.F. & Cavalieri, D. (2003). Fundamentals of cDNA Microarray Data Analysis. TRENDS GENET 19, 649-659.

Lian, W. et al. (2008). Genome-wide transcriptome analysis reveals that a pleiotropic antibiotic regulator, AfsS, modulates nutritional stress response in Streptomyces coelicolor A3(2). BMC Genomics 9, 56.

Liao, L., Noble, W.S. (2003). Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J. Comput. Biol 10, 857-868.

Margulies, M. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.

Martinez, R., Pasquier, N. & Pasquier, C. GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. Bioinformatics 24, 2643-2644 (2008).

Moreau, Y. et al. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet 19, 570-577.

Mootha, V.K. et al. (2003). Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proceedings of the National Academy of Sciences of the United States of America 100, 605-610.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol 247, 536-540.

Newman, J.C., Weiner, A.M. (2005). L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol 6, R81.

O'Neill, M. & Ryan, C. (2003). Grammatical evolution: evolutionary automatic programming in an arbitrary language. Springer.

O'Neill, M. et al. (2005). A grammatical evolution approach to eukaryotic promoter recognition, Bioinformatics Inform Workshop and Symposium, Dublin, Ireland.

O'Neill, M. & Brabazon, A. (2006). Grammatical Swarm: The generation of programs by social programming. Natural Computing 5, 443-462.

O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. J R Soc Med 100, 579-582.

Pavey S. et al. Hayward. (2004). Microarray expression profiling in melanoma reveals a BRAF mutation signature. Oncogene, 23(23):4060-4067, May 2004.

Piatetsky-Shapiro, G. & Tamayo, P. (2003). Microarray data mining: facing the challenges. SIGKDD Explor. Newsl. 5, 1-5.

Pomeroy, S.L.,Tamayo,P.,Gaasenbeek, M.,Sturla, L.M.,Angelo, M.,McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415(6 870) pp 436–442.

Ray, H.N., Mootha, V.K., Boxwala, A. A. (2003). Building an Application Framework for Integrative Genomics. AMIA Annu Symp Proc 2003, 981-981.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13, 163.

Ren, B. et al. (2000). Genome-Wide Location and Function of DNA Binding Proteins. Science 290, 2306-2309.

Richter, B.G., Sexton, D.P. (2009). Managing and Analyzing Next-Generation Sequence Data. PLoS Comput Biol 5, e1000369.

Rhodes, D.R. et al. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 62, 4427-4433.

Rhodes, D.R. et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 6, 1-6.

Rost, B. (2001). Review: Protein Secondary Structure Prediction Continues to Rise. J. STRUCT. BIOL 134, 204-218.

Raychaudhuri, S., Stuart, J.M., Altman, R.B. (2000). Principal Components Analysis To Summarize Microarray Experiments: Application To Sporulation Time Series. In Pacific Symposium on Biocomputing 5, 452-463.

Saigo, H., Vert, J., Ueda, N., Akutsu (2004). T. Protein homology detection using string alignment kernels. Bioinformatics 20, 1682-1689.

Segal, E. et al. (2004). A module map showing conditional activity of expression modules in cancer. Nat. Genet 36, 1090-1098.

Sansone, S. et al. (2008). The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?". OMICS 12, 143-149.

Staden, R. (1977). Sequence data handling by computer. Nucleic Acids Res 4, 4037-4051.

Stuart, J.M. et al. (2003). A gene-coexpression network for global discovery of conserved genetic modules. SCIENCE 302, 249-255.

Smith, B. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol 25, 1251-1255.

Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology 147, 195-197.

Thieblemont C. et al. (2004). Smal lymphocytic lymphoma, marginal zone B-cell lymphoma, and mantle cell lymphoma exhibit distinct gene-expression profiles allowing molecular diagnosis. Blood 2004; 103: 2727-2737.

Tamayo, P. et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America 96, 2907-2912.

Tavazoie, S. et al. (1999). Systematic determination of genetic network architecture. Nat. Genet 22, 281-285.

Taylor, C.F. et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat. Biotechnol 26, 889-896.

Taylor, W.R. (1986). The classification of amino acid conservation. J. Theor. Biol 119, 205-218.

Troyanskaya, O. et al. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520-525.

Tusher, V.G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U.S.A 98, 5116-5121.

Vapnik, V.N. (2000). The nature of statistical learning theory, Springer.

Wang, J., T. H. Bø, I. Jonassen, O. Myklebost, Hovig, E. (2003). Tumor classification and marker gene prediction by feature selection and fuzzy cmeans clustering using microarray data, BMC Bioinformatics 4, 60.

Wang, J. et al. (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. Bioinformatics 20, 3166-3178.

Watson, J.D. & Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737-738.

Wren, J.D., Johnson, D. & Gruenwald, L. (2005). Automating Genomic Data Mining via a Sequence-based Matrix Format and Associative Rule Set. BMC Bioinformatics 6, S2-S2.

Workman, C. et al. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biology 3, research0048.1-research0048.16.

Wu, J. et al. (2009). Web-based interrogation of gene expression signatures using EXALT. BMC Bioinformatics 10, 420.

Yan, L., Dodier, R., Mozer, M.C., Wolniewicz, R. (2003). Optimizing Classifier Performance Via the Wilcoxon-Mann-Whitney Statistic.

Yang, Y.H. et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucl. Acids Res. 30, e15.

Yeang,C.H.,Ramaswamy,S.,Tamayo,P.,Mukherjee,S.,Rifkin,R.M.,Angelo,M., Reich, M., Lander, E., Mesirov, J., Golub, T. (2001). Molecular classification of multiple tumor types. Bioinformatics 17 Suppl 1 316–322.

Zemla, A., Venclovas, C., Fidelis, K., Rost, B. (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 34, 220-223.

Zhou, X.J. et al. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. Nat. Biotechnol 23, 238-243.

Zhu, J. et al. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40, 854-861.

Curriculum Vitae

Organisation du Curriculum Vitae

1 Etat civil et situation administrative
2 Formation
3 Parcours
4 Thèmes de Recherche
5 Publications
6 Encadrement
7 Responsabilités scientifiques
8 Rayonnement scientifique
9 Responsabilités administratives
10 Enseignement

1. Etat civil et situation administrative

Nom	Ramstein
Prénom	Gérard
Date de naissance	05/10/1959 à Cernay (68)
Nationalité	Française
Situation familiale	marié
Grade	3014 - MCF CN
Etablissement d'affectation Section CNU :	Ecole polytechnique de l'université de Nantes 27.0 Informatique
Date de titularisation	01/10/1992
Date d'affectation	01/10/1990
Adresse administrative	Polytech'Nantes Site de la Chantrerie - Rue Christian Pauc BP 50609 – 44 306 Nantes Cedex 3
Téléphone	02 40 68 30 16
Adresse électronique	gerard.ramstein@univ-nantes.fr

2. Formation

1989-1990

Post-doctorat au Groupement Scientifique de Télédétection Spatiale de Strasbourg (GSTS) de Strasbourg, bourse du CNES.

1986-1989

Thèse de doctorat de l'université Louis Pasteur de Strasbourg 1, spécialité Informatique.
Structures spatiales irrégulières dans les images de Télédétection. Applications de la notion de dimension fractale.

Thèse obtenue avec mention Très Honorable, soutenue le 13 septembre 1989.
Composition du jury :

M. R. Armbruster	Président
M. F. Becker	Directeur de thèse
M. J. Francon	Rapporteur
M. R. Jeansoulin	Rapporteur
M. M. Raffy	Directeur de thèse

1985-1986
DEA Traitements d'images – Traitements graphiques, université Louis Pasteur, Strasbourg 1.

1980-1982
Diplôme d'ingénieur de l'Ecole Nationale Supérieur des Arts et Industries de Strasbourg, option topographie.

3. Parcours

depuis 1990
Maître de Conférences en informatique à l'Ecole polytechnique de l'université de Nantes.

- **1990 – 1993**
  Membre du Laboratoire d'Analyse et de Traitement d'Images (LATI), dirigé par D. Barba, ce laboratoire est actuellement intégré à l'IRCCyN (Institut de Recherche en Communications et Cybernétique de Nantes).
- **1994 – 1999**
  Membre du laboratoire Systèmes Electroniques et Informatiques (SEI), dirigé par J. Saillard, ce laboratoire est actuellement intégré à l'IRCCyN.
- **depuis 2000**
  Membre du Laboratoire d'Informatique de Nantes-Atlantique (UMR 6241 LINA), équipe COnnaissances et Décision (COD).

4. Thèmes de recherche

Le traitement d'images étant ma discipline de recherche d'origine, j'ai été accueilli dès ma prise de fonction en 1990 dans une équipe spécialisée dans cette thématique (Laboratoire d'Analyse et de Traitement d'Images (LATI, aujourd'hui dissolu). Mon intérêt s'est porté dès mon intégration vers les applications biologiques ; mes premiers travaux avaient porté sur l'analyse des chromosomes par traitement d'images. Il s'agissait de développer une boîte à outils pour analyser le génome d'un individu et découvrir d'éventuelles aberrations chromosomiques. En collaboration avec Maurice Bernadet, également membre du LATI, nous avons ensuite piloté cet outil par un système à base de règles pour automatiser la classification des chromosomes. En 1994, le laboratoire a fusionné avec le S2HF (dirigé alors par J. Saillard) pour former le SEI (Systèmes Electroniques et Informatiques). Suite à cette fusion, un remaniement des thèmes a été défini. Mon thème de recherche a été abandonné pour des raisons stratégiques : le champ était alors couvert par de nombreuses publications et toute innovation demandait un investissement jugé trop important. J'ai alors intégré un axe transversal portant sur les architectures matérielles dédiées au traitement d'images. Deux ans plus tard, le SEI fusionnait avec un nouveau laboratoire, les équipes respectives intégrant l'Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN). L'axe transversal n'avait plus lieu d'être dans cette nouvelle structure. Comme les activités de l'équipe IVC (Images et Vidéo Communication) étaient principalement orientées Traitement du Signal (notamment en compression d'images), j'ai pris la décision de rejoindre une équipe dont la recherche était clairement identifiée comme appartenant à l'informatique, en cohérence avec ma discipline de prédilection. En accord avec le directeur de l'équipe d'alors, Henri Briand, j'ai pu reprendre mon champ d'application

premier, à savoir la biologie, en développant des techniques de *data mining* pour les données biomédicales.
La suite de cette section reprend les différents thèmes de recherche auxquels je me suis confronté depuis ma nomination en omettant la période 1994-1999 consacrée à la modélisation des architectures spécifiques pour l'analyse d'images.

4.1 Thème de recherche dans le cadre du LATI (1990-1993) Classification automatique de chromosomes par analyse d'images.

Mon travail consistait à définir un système automatique de constitution d'un caryotype à partir d'une image prise au microscope. Le caryotype est la cartographie complète des chromosomes établie selon un classement conventionnel (incluant notamment le regroupement par paires). Le système développé comprenait un ensemble d'outils dédiés, essentiellement des procédures de traitement d'images, de reconnaissance de formes et de classification. L'analyse du caryotype est un problème d'une grande complexité : malgré des techniques de préparation permettant un étalement des chromosomes, ceux-ci peuvent se chevaucher et faire croire à des anomalies de recombinaison chromosomique. Une technique de coloration fait apparaître une alternance de bandes sombres et claires. La structure de ces bandes permet de caractériser les chromosomes, ce qui m'a permis de développer des méthodes d'identification du caryotype d'un patient. La chaîne de traitement mise au point incluait la segmentation de l'image métaphasique, l'extraction des caractéristiques morphométriques, l'analyse du profil chromosomique et sa classification. Une nouvelle méthode de classification a été élaborée, basée sur une reconnaissance de formes structurelles avec optimisation par recuit simulé. Une approche par réseaux neuronaux a également été développée. Enfin, un système à base de règles a été conçu. Les règles ont été définies avec l'aide de cytogénéticiens pour déterminer correctement les classes chromosomiques. De nombreuses heuristiques guident en effet l'analyse, l'expert résolvant les confusions interclasses par des règles de discrimination dépendant des classes concernées.

4.2 Thème de recherche dans le cadre du LINA/COD (depuis 2000) Application des techniques de *data mining* aux données biologiques

Mon thème de recherche actuel concerne l'application des techniques d'extraction de connaissances à partir de données (ECD) en biologie. L'essor des techniques de biologie moléculaire à partir des années 50 a profondément transformé l'activité des chercheurs de ce domaine. Face à l'accroissement considérable du volume de données lié notamment aux techniques de séquençage des génomes, la nécessité de développer de nouvelles méthodes de travail s'est révélée, donnant naissance à un néologisme : la bio-informatique, discipline au croisement de la biologie et de l'informatique (en vérité il faut également y inclure les mathématiques pour la résolution des problèmes de modélisation et de statistiques).

Au cœur de chacune de nos cellules se trouve notre patrimoine génétique : l'ensemble de nos gènes ou génome. Ces gènes peuvent être considérés comme le « plan de construction » permettant à nos cellules de produire les protéines, l'un des matériaux de base pour la construction de notre organisme. Ils sont dits exprimés lorsque la cellule enclenche la transcription du « plan » conservé dans le noyau de la cellule pour fabriquer une nouvelle protéine : on appelle transcriptome l'ensemble des transcriptions d'un organisme. La séquence primaire d'une protéine, déterminée par la succession des acides aminés qui la composent, définit sa forme et donc sa fonction dans l'organisme.

Mon thème de recherche s’est confronté à deux axes fondamentaux dans cette récente discipline, à savoir la recherche d’homologues distants dans des familles de protéines et l’analyse du transcriptome.

Recherche d’homologues distants à partir de séquences protéiques

Cette problématique consiste à découvrir de nouveaux membres d’une famille de protéines. Un effort considérable a été entrepris afin de classer les protéines connues en familles et en superfamilles. Les membres de ces classes partagent généralement une fonction biologique précise. Identifier l’appartenance d’une protéine permet donc d’inférer son rôle dans l’organisme. La famille d’intérêt, étudiée en collaboration avec Yannick Jacques (responsable de l’équipe *Cytokines et Récepteurs en Immuno-Cancérologie*, unité Inserm U892, Centre de Recherche en Cancérologie Nantes-Angers), est la superfamille des cytokines. Ces protéines sont impliquées dans le système immunitaire et constituent une des plus grandes et des plus hétérogènes des familles de protéines.
Ma contribution a consisté à développer des classifieurs capables de discriminer une famille de protéines particulière. La technique des Séparateurs à Vastes Marges (SVM) a été retenue, cette méthode ayant donné les résultats les plus prometteurs pour ce type d’application. Parmi les travaux menés, une méthode originale de classification a été conçue et une technique originale de fusion de classifieurs a été définie, basée sur les essaims grammaticaux.

Analyse des données du transcriptome

La technologie des puces à ADN permet d'analyser l'expression simultanée de milliers de gènes. L’étude du transcriptome représente un enjeu considérable, tant du point de vue de la compréhension des mécanismes du vivant que des applications cliniques et pharmacologiques. L’étude du transcriptome permet de répondre à de nombreuses questions biologiques. Je me suis notamment intéressé à la co-expression des gènes : il a été montré que des gènes ayant des mécanismes communs d’expression partagent souvent une fonction biologique et que l’analyse de la co-expression est une voie d’accès à l’inférence de réseaux de régulation, permettant ainsi de mieux comprendre le fonctionnement cellulaire. J’ai proposé d’appliquer la méthode d'analyse implicative des règles d'association sur les données du transcriptome. J'ai pu montrer que cette approche, en s'intéressant à des similarités locales dans les rangs, est bien plus fine que les mesures globales et qu'elle permet de mettre en évidence des associations réelles entre des groupes d'observations présentant le même phénotype. J'ai également utilisé cette mesure pour développer une méthode originale de sélection de gènes discriminants et cette méthode a été appliquée pour établir des règles de classification.

Grâce à une collaboration étroite avec l'équipe INSERM U915 dirigée par Rémi Houlgatte (et avant lui Jean Léger), j'ai pu aider à la création des outils MADTOOLS, une suite logicielle dédiée à l'analyse de puces à ADN. La base MADGene (qui fait partie de cette suite) associe à chaque sonde un identifiant unique et son moteur de recherche est capable d'identifier une sonde quelles que soient la nomenclature utilisée et son espèce. Cet outil intégratif a permis la définition d'une base de données dédiée aux pathologies musculaires (MADMuscle).

5. Publications

Chapitre d'ouvrages collectifs

G. Ramstein
Gravitational singularities in particle swarms. Application to the discovery of gene expression patterns in DNA microarrays, Evolutionary Computation, Evolutionary Computation, Wellington Pinheiro dos Santos (Ed.), pp 405-416, InTech, 2010.

G. Ramstein, N. Beaume, and Y. Jacques
Detection of Remote Protein Homologs using Social Programming
Foundations of Computational Intelligence vol. 4, Bio-Inspired Data Mining, ed. Springer Berlin / Heidelberg, pp. 277-296, 2009.

G. Ramstein
Une méthode implicative pour l'analyse de données d'expression de gènes, Analyse Statistique Implicative, Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd, invités R. Gras, J. C Régnier, F. Guillet, Cépaduès Ed. Toulouse, pp 455-469, 2009.

G. Ramstein
Statistical Implicative Analysis of DNA microarrays.
Statistical Implicative Analysis - Theory and Applications Series: Studies in Computational Intelligence , vol. 127, pp 205-226. Springer. Gras, R.; Suzuki, E.; Guillet, F.; Spagnolo, F. (Eds.), 2008.

Revues internationales

Daniel Baron, Armelle Magot, Gerard Ramstein, Marja Steenman, Guillemette Fayet, Catherine Chevalier, Philippe Jourdon, Remi Houlgatte, Frederique Savagner, Yann Pereon
Immune Response and Mitochondrial Metabolism Are Commonly Deregulated in DMD and Aging Skeletal Muscle. **PLoS ONE** (I.F. 4,4), 6(11), 2011.

Daniel Baron, Audrey Bihouee, Raluca Teusan, Emeric Dubois, Frederique Savagner, Marja Steenman, Remi Houlgatte, Gerard Ramstein.
MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets. **Bioinformatics** (I.F. 4,9), 27, 725 –726, 2011.

Daniel Baron, Emeric Dubois, Audrey Bihouee, Raluca Teusan, Marja Steenman, Philippe Jourdon, Armelle Magot, Yann Pereon, Reiner Veitia, Frederique Savagner, Gerard Ramstein, Remi Houlgatte
Meta-analysis of muscle transcriptome data using the MADMuscle database reveals biologically relevant gene patterns. **BMC Genomics** (I.F. 3,8), 12, 113, 2011.

Lamirault, G.; Le Meur, N.; Roussel, J.C.; Baron, D.; Bihouée, A.; Guisle, I.; Raharijaona, M.; Ramstein, G.;Teusan, R.; Chevalier, C.; Gueffet, J.P.; Trochu, J.N.; Léger, J.; Houlgatte, R.; Steenman, M.
Molecular risk stratification in advanced heart failure patients. **Journal of Cellular and Molecular Medicine** (I.F. 5,3), 14(6B), 1443-52, 2010**.**

Le Meur N., Lamirault G., Bihouée A., Steenman M., Bédrine-Ferran H., Teusan R., Ramstein G., Léger J.J.
A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values. Importance of replication, **Nucleic Acids Research** (I.F. 7,5), vol. 32, n° 18, pp. 5349-5358, Oct. 2004.

G. Ramstein, M. Bernadet
Chromosome: a knowledge-based system for chromosome classification, **Technology and Health Care**, vol. 1, n° 2, pp. 171-180, November 1993.

Ramstein, G., and M. Raffy
Algorithme d'analyse fractale de contours en télédétection et applications **International Journal of Remote Sensing**, vol. 11, n° 2, pp. 191-208, 1990.

Ramstein, G., and M. Raffy
Analysis of the structure of remotely-sensed images,
International Journal of Remote Sensing, vol. 10, pp. 1049-1073, 1989.

Revues nationales

G. Ramstein, M. Raffy
Restauration d'images de télédétection par variogramme, **Traitement du Signal**, 1988, vol. 5, n° 4, pp 263-272, 1988.

Conférences et workshops internationaux avec actes publiés

H.T. Nguyen, P. Leray, G. Ramstein
Multiple hypothesis testing and quasi essential graph for comparing two sets of Bayesian networks, in proceedings of the 15th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), pp 176-185, Lecture Notes in Artificial Intelligence LNAI, Springer-Verlag, , Kaiserslautern, Germany , September, 12-14, 2011

H.T. Nguyen, P. Leray, G. Ramstein
Summarizing and visualizing a set of Bayesian networks with quasi essential graphs, in proceedings of the 14th International Conference of the Applied Stochastic Models and Data Analysis (ASMDA), Roma, Italy, June, 7-10, 2011.

G. Ramstein, N. Beaume, Y. Jacques
A Grammatical Swarm for Protein Classification, in WCCI 2008 Proceedings (IEEE World Congress on Computational Intelligence, including 2008 IEEE Congress on Evolutionary Computation, pp. 2561-2568, Hong Kong, China, June 1-6, 2008.

N. Beaume, G. Ramstein, Y. Jacques
An expert-based approach for the identification of remote homologs, in Fith Int. Worshop on Computational Systems Biology, WCSB08, pp. 17-20, Leipzig, Germany, June 11-13 2008.

J. Lorec, G. Ramstein, Y. Jacques
Identifying heterogeneous and complex named entities in biology texts using controlled dictionaries, in ECML/PKDD Workshop on Data and Text Mining for Integrative Biology, pp. 28-39, Berlin, September 18-22, 2006.

J. Mikolajczak, G. Ramstein, Y. Jacques
SVM-based classification of distant proteins using hierarchical motifs, in Fifth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'04), Springer, pp. 25-30, Exeter, UK, 25-27 August, 2004.

Raluca Teusan, Audrey Bihouee, Nolwenn Le Meur, Gérard Ramstein, Jean Leger
MADTools: management tool for the mining of microarray data. 'Details on MADSENSE, a gene comprehension support system', in European Conference on Computational Biology, ECCB, pp. 391-392, Paris, 27-30 September 2003.

G. Ramstein, P. Bunelle, Y. Jacques.
Discovery of ambiguous patterns in sequences. Application to bioinformatics, in 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'2000), pp 581 – 586, September 2000.

G. Ramstein, F. Bouchard, P. Bakowski, V. Stachetti
Aspects in Modelling the Data Path of a Real Complex DSP Component, in European Design & Test Conference, ED&TC'97, pp. 128-134, Paris, 17-20 Mars 1997.

P. Collet, G. Ramstein, F. Megel
SAGAPA or How to Design Minimal DSP-based Systems for Image Processing Applications, in DSP'97, pp. 392-398, Paris, 17-19 Septembre 1997.

P. Collet, G. Ramstein, B. Longuet
A Computer Aided Architecture Design Tool Aimed at Image Processing Applications, in Int. Conference on High-Performance Computing, HIPC'97, pp. 242-248, Bangalore, Inde, 18-21 December 1997.

G. Ramstein, O. Deforges, P. Bakowski
A Design Tool For The Specification And The Simulation Of Array Processors Architectures. Application To Image Processing: The Extraction Of Regions Of Interests, in Application Specific Array Processors ASAP'95, pp. 296-344, Strasbourg, 24-26 Juillet 1995.

G. Ramstein, P.Y. Le Meur, D. Barba
A structural method for the chromosome classification, in Europ. Conf. on Engineering and Medecine 93, pp. 132-34, Stuttgart, 25-28 April 1993.

G. Ramstein, M. Bernadet, A. Kangoud, D. Barba
Chromosome : A Rule-Based Image Analysis System For Chromosome classification, in 14th Conference IEEE/EMBS, pp. 926-927, Paris, October 92.

D. Barba, X.Qiu, G. Ramstein
Image restoration for chromosome classification with band information, in 9th Inter. Conf. on Digital Signal Processing, pp. 186-202, Florence, Italie, 4-6 Septembre 1991.

G. Ramstein
An interpolation method for stochastic models, in European Computer Graphics Conference, pp 353-364, Montreux, Suisse, 3-7 septembre 1990.

M. Raffy, G. Ramstein, F. Becker
Fractal analysis of a relief and its remotely sensed image, in Second Image Symposium. Int. Electronic Image Week, pp. 391-392, Nice, Avril 1986.

Conférences internationales avec comité de sélection et actes à diffusion restreinte

H.T. Nguyen, G. Ramstein, P. Leray, Y. Jacques
Differential study of the cytokine network in the immune system by the evolutionary algorithm based on the bayesian network, In proceedings of the 2nd Asian Conference on Intelligent Information and Database Systems (ACIIDS), Doctoral Colloquium, Hue City, Vietnam , March, 24-26, 2010.

Nolwenn Le Meur, Audrey Bihouee, Raluca Teusan, Gérard Ramstein, Jean Leger
MADTools : management tool for the mining of microarray data. 'Details on MADSCAN, a tool for processing chip data, in Cambridge Healthtech Institute's –Microarray Data Analysis Baltimore (US), 21-23 September 2003.

G. Ramstein, P. Bakowski, Y. Nadreau
VHDLP: specification and evaluation of performance constraints for design tuning.
Forum on Design Languages, FDL'99, Nantes, 30 Août - 3 Septembre 1999.

G. Ramstein, P. Bakowski
A front end VHDL environment for By-Default Specification, Simulation and Synthesis of Array Processors Architectures. VHDL-Forum for CAD in Europe, Spring94 Meeting, Lago di Como, Italie, 17-20 April 1994.

M. Bernadet, H. Benali, G. Ramstein
Three multiagent architectures for medical image processing. Congres EXPERSYS 92, Paris, 21-22 Octobre 1992.

Conférences nationales avec actes publiés

G. Ramstein
Une méthode implicative pour l'analyse de données d'expression de gènes, Quatrième rencontre internationale A.S.I, Analyse Statistique Implicative, pp 123-134, Ed. R.Gras, Castellon, Espagne, 18-21 Octobre 2007.

J. Lorec, G. Ramstein, Y. Jacques
Extraction et identification d'entités complexes à partir de textes biomédicaux. CEPADUES éditions, Revue des Nouvelles Technologies de l'Information RNTI-E-6, vol. 1, pp 223-228, 2006.

J. Mikolajczak, G. Ramstein, Y. Jacques
Détection de faibles homologies de proteines par machines à vecteurs de support, RNTI Classification et fouille de données, publication RNTI-C-1, Cépadues-édition, 11èmes Rencontres de la Société Francophone de Classification, pp. 89-100, Bordeaux, 8-10 septembre 2004.

J. Mikolajczak, G. Ramstein, Y. Jacques
Caractérisation de signatures complexes dans des familles de protéines distantes, Colloque Extraction et Gestion des Connaissances (EGC'04), éditeur CEPADUES-EDITIONS, publication RNTI,vol. 2, pp 317-328, 20-23 janvier 2004.

H. Delalin, J. Léger, G. Ramstein
Analyse de l'expression de gènes à partir de puces à ADN, Journées Francophones d'Extraction et de Gestion des Connaissances (EGC'02), vol. 4, n° 1, pp. 421-421, Hermès, Montpellier, 21-23 janvier 2002.

G. Ramstein, P. Bunelle and Y. Jacques.
Algorithmes pour l'analyse de séquences biologiques. Extraction des Connaissances et apprentissage, Hermès, pp. 141-147, vol. 1, n°1-2, Nantes, 17-19 janvier 2001.

M. Bernadet, G. Ramstein, H. Benali
Massage: un générateur de systèmes multi-agents distribués, spécificités et exemples de réalisation. Intelligence Artificielle Distribuée et Systèmes Multi-Agents, pp. 86-98, Toulouse, 7-8 Avril 1993.

D. Barba, X.Qiu, G. Ramstein
La classification automatique des chromosomes par analyse d'images, 8ème Congres R.F.I.A, Lyon, pp. 545-549, 25-29 novembre 1991.

D. Barba, X.Qiu, G. Ramstein
Performances comparées de méthodes de classification de chromosomes, 3ème Colloque GRETSI, pp. 1090-1093, Juan-les-Pins, Septembre 1991.

Conférences et ateliers nationaux avec comité de sélection et actes à diffusion restreinte

H.T. Nguyen, G. Ramstein, P. Leray, Y. Jacques
Reconstruction de réseaux de régulations génétiques par l'approche évolutionnaire sur les réseaux Bayésiens, in proceedings of MODGRAPH, Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Nantes, June, 9-11, 2009.

G. Ramstein
Une méthode implicative pour l'analyse de données d'expression de gènes, Quatrième rencontre internationale A.S.I, Analyse Statistique Implicative, Ed. R.GRas, Castellon, Espagne, 18-21 Octobre 2007.

N. Beaume, G. Ramstein, Y. Jacques
Searching for remote homologs : a combined approach based on SVM and experts, 7èmes Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), Bordeaux, 5-7 juillet 2006.

N. Beaume, J. Mikolajczak, G. Ramstein, Y. Jacques
Recherche de nouveaux membres de la superfamille des cytokines par Séparateurs à Vastes Marges. 6ème Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), Lyon, 6-8 Juillet 2005.

G. Ramstein
Vers une analyse implicative des données issues de puces à ADN, Troisièmes rencontres internationales A.S.I, Analyse Statistique Implicative, Palerme, 6-8 Octobre 2005.

J. Mikolajczak, G. Ramstein, Y. Jacques
Classification de protéines distantes par motifs hiérarchiques, 5èmes Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), Montréal , 28 - 30 juin 2004.

H.Delalin and J. Leger and G.Ramstein
Extraction de patterns flous sur des données du transcriptome. Atelier puces d'Extraction et Gestion des Connaissances 2003.

G. Ramstein, M.Bernadet, A. Kangoud, D. Barba
A knowledge-based system for chromosome classification, Congres de Cytométrie en Image, Montpellier, 2-4 Decembre 1992.

G. Ramstein, M. Raffy
Interprétation morphologique de structures. Application à la reconnaissance du couvert forestier par télédétection aéroportée. Machines et Réseaux Intelligents, Paris, mai 1987.

G. Ramstein, M. Raffy. Caractérisation multispectrale de la structure spatiale des scènes télédétectées. Signatures Spectrales d'Objets en Télédétection. Quatrième Colloque Intern., Aussois, Janvier 1988.

Rapports

G. Ramstein, O. Deforges
Conception et Validation d'une Architecture Matérielle de Traitement en vue d'une Application de Tri Postal. Contrat SRTP (Service de Recherche Technique de la Poste), rapport no 94.0367.7680.CS1.00, Décembre 1994.

6. Encadrements

6.1 Encadrements de thèses

Hoai Tuong Nguyen, thèse co-encadrée avec Philippe Leray (Professeur des Universités, LINA) et Yannick Jacques (Directeur de recherche, INSERM U463), *Réseaux bayésiens et apprentissage ensembliste pour l'étude différentielle de réseaux de régulation génétique*, soutenance prévue le 27 janvier 2012.

Julien Lorec, thèse co-encadrée avec Yannick Jacques de l'équipe INSERM U463, *Extraction de réseaux de régulation à partir de textes biomédicaux*, Université de Nantes, thèse soutenue le 2 octobre 2008.

Nicolas Beaume, thèse co-encadrée avec Yannick Jacques de l'équipe INSERM U463, *Une approche multicritère pour la recherche d'homologues distants,* Université de Nantes, thèse soutenue le 27 juin 2008.

Jérôme Mikolajczak, thèse co-encadrée avec Yannick Jacques de l'équipe INSERM U463, *Caractérisation d'une famille de protéines à partir d'informations structurales multiples*, Université de Nantes, thèse soutenue le 4 mai 2005.

Pascal Collet, thèse co-encadrée avec Dominique Barba de l'IRCCyN, *Méthodes de conception d'une architecture optimisée de traitement pour des applications d'images,* Université de Nantes, thèse soutenue le 28 novembre 1997.

Olivier Deforges, thèse co-encadrée avec Dominique Barba de l'IRCCyN, *Segmentation robuste d'images de documents par une approche multirésolution. Conception et validation d'une architecture parallèle dédiée*, Université de Nantes, thèse soutenue en septembre 1995.

6.2 Encadrements de masters

Kévin Panhalleux

Méta-analyse de données de puces à ADN.

Master professionnel de Bio-informatique de Nantes, 2010.

Florent Dumont

Analyse d'une méthode de clusterisation de données de puces à ADN.

Master professionnel de Bio-informatique de Nantes, 2009.

Coline Thomas

Identification d'homologues distants. Application à la famille des cytokines.

Master professionnel de Bio-informatique de Nantes, 2008.

Hela Memni Rekik

Extraction de Signature de Gènes à Partir d'Etudes Multiples de Puces à ADN.

Master Systèmes d'Aide à la Décision, 2008

Hoai-Tuong Nguyen

Méta-analyse de données de puces à ADN, master Extraction de Connaissances à partir de Données, 2008.

Hugues Delalin.

Analyse de l'expression de gènes à partir de puces à ADN.

Master Extraction de Connaissances à partir de Données, 2002.

Frédérique Pesquet

Projection d'algorithmes de traitement d'images sur des architectures massivement parallèles, DEA, septembre 1994.

Jean-Luc Panier

Classification automatique de chromosomes par réseau neuromimétique, DEA, septembre 1993.

Pierre-Yves Lemeur

Reconnaissance de chromosomes par analyse structurelle du profil chromosomique, DEA, septembre 1992.

6.3 Encadrement de Mémoire d'ingénieur CNAM

Pascal Bunelle, Mémoire d'ingénieur CNAM, en collaboration avec Yannick Jacques de l'équipe Inserm U463, Application biologique et Data-mining : Recherche de motifs conservés dans une famille de protéines : les récepteurs de cytokine, Soutenu le 08/03/2001

7. Responsabilités scientifiques

Je présente dans cette section les responsabilités que j'exerce dans le cadre de mes activités scientifiques. Je suis responsable de la plate-forme bioinformatique de Nantes, de la co-animation de l'axe bio-informatique de Biogenouest et je participe à deux conseils scientifiques.

7.1 Responsabilité de la plate-forme bio-informatique de Nantes

La plate-forme de bio-informatique de Nantes est une composante de l'Institut fédératif de recherche thérapeutique de Nantes (IFR 26 - Inserm/MESR/Université) . Elle a pour mission de concevoir et de développer des outils innovants en recherche bio-informatique, ainsi que de proposer un ensemble de services pour l'analyse bio-informatique. L'activité de recherche a conduit au développement d'un système de biologie intégrative comportant plusieurs outils disponibles sur internet (www.madtools.org) et a donné lieu à de nombreuses publications auxquelles j'ai participé. L'activité de service consiste à aider les utilisateurs dans leur analyse, sous la forme d'une prestation de service, des données brutes à l'interprétation biologique des résultats, ou sous la forme d'une aide intégrant formation et expertise. La plate-forme emploie trois ingénieurs d'étude. Elle fait partie de l'axe bio-informatique *de Biogenouest*.

7.2 Conseils scientifiques

Membre du Conseil Scientifique de Biogenouest

Le Conseil Scientifique se réunit une fois par mois pour définir la politique scientifique de Biogenouest.

Membre du Conseil Scientifique de la startup GeneNext depuis 2009

Le projet GeneNext a été lauréat en 2009 du Concours national d'aide à la création d'entreprises de technologies innovantes organisé par le Ministère de l'enseignement supérieur et de la recherche. GeneNext est dédié à la biologie intégrative des pathologies thyroïdiennes et cardiaques. L'objectif de GeneNext est d'intégrer les technologies de la génomique fonctionnelle pour accélérer la découverte de biomarqueurs et de cibles thérapeutiques dans les pathologies thyroïdiennes et cardiaques.

8. Rayonnement scientifique

8.1 Organisation de colloques et d'ateliers

- Coordonnateur du pole Communication du Comité d'Organisation des *Journées Ouvertes en Biologie, Informatique et Mathématiques 2009* (JOBIM), organisées du 9 au 11 juin à Nantes.

- Co-organisateur de la journée satellite *Modèles graphiques probabilistes pour l'intégration de données hétérogènes et la découverte de modèles causaux en biologie* (MODGRAPH), le 8 juin 2009.

- Co-organisateur de l'atelier Extraction et gestion de connaissances appliquées aux données biologiques, Lille, le 17 janvier 2006.

- Participation à l'organisation du colloque *Extraction et Gestion des Connaissances*, Nantes, 17-19 janvier 2001.

- Participation à l'organisation du *colloque VHDL-FORUM for CAD in EUROPE*, 24-27 avril, IRESTE-IHT, Nantes 1995.

8.2 Jury de thèse

Outre les participations aux jurys des thèses dont j'ai assuré le co-encadrement, j'ai été examinateur du jury de thèse de Raharijaona Mahatsangy :
De la génomique fonctionnelle vers la génomique intégrative de pathologies humaines.
Université de Nantes, 15 octobre 2009.

8.3 Participations à des réseaux et à des groupes de travail scientifiques

- Membre depuis 2008 du réseau de BioInformatique Ligérienne (BIL). L'objectif de ce réseau est de fédérer et de promouvoir la recherche et l'enseignement en Génomique-Bioinformatique dans la région Pays de Loire. Le réseau comporte dix équipes dont l'équipe COD du LINA. Certaines de ces équipes sont impliquées dans la production de données génomiques et développent des approches fonctionnelles à l'échelle du génome. D'autres développent des méthodes innovantes d'analyse. Un groupe de travail collaboratif a été mis en place pour faire émerger des travaux multidisciplinaires. Ce réseau a été doté d'un financement par la Région de l'ordre de 1,2 million €. Six thèses ont ainsi été lancées en 2008, en co-encadrement d'un informaticien et d'un biologiste.

- Groupe de travail sur le transcriptome et le data mining, Action Spécifique Gafodonnées (2002).

- Recherche et Extraction de Motifs pour l'Analyse Génomique, Action de recherche coopérative INRIA (2000).

- GDR TDSI 6: Architecture pour le Signal et l'Image (1998)

- PRC-GDR Architectures nouvelles de Machines: Architectures spécialisées pour la reconnaissance de formes et le traitement du signal (1997).

8.4 Transfert technologique

Collaboration avec le Service de Recherche Technique de la Poste à Nantes en 1994

Cette collaboration portait sur la reconnaissance du bloc adresse destinataire sur des colis postaux (contrat SRTP, rapport no 94.0367.7680.CS1.00, décembre 1994).

9. Responsabilités administratives

Chargé de mission sur les Technologies de l'Information et de la Communication pour l'Education (TICE) depuis 2002

La direction de Polytech'Nantes m'a chargé de favoriser l'utilisation des TICE, de coordonner les activités autour des TICE sur l'ensemble des départements de l'école ainsi que de représenter celle-ci auprès des instances concernées. L'innovation numérique est une priorité pour les Pays de Loire, comme le montre l'aménagement régional en infrastructure très haut-débit (réseau *Gigalis*) et le développement de l'Université Numérique en Région (UNR).
J'ai aidé à la mise en place des ressources pédagogiques numériques, en collaboration avec le service informatique de l'université. Après concertation avec les enseignants de l'école, j'ai demandé à ce que soient adaptés les outils à nos besoins propres, notamment en ce qui concerne la création des comptes utilisateurs et leur associations aux espaces pédagogiques.

Membre du Conseil de Direction (1990-1995)

J'ai participé dans les premières années de ma fonction au Conseil de Direction de l'école. Ce conseil est l'organe de direction de l'école. Il définit la politique de l'établissement et soumettait ses projets au Conseil d'Administration.

Responsable de la discipline Informatique pour les élèves-ingénieurs de Polytech'Nantes, spécialité ETN, 3ème année depuis 1995

Je coordonne le contenu des deux modules pédagogiques en informatique pour la troisième année du département ETN (promotion de 72 étudiants).

Membre du groupe TICE de la Conférence des Grandes Ecoles du Pays de Loire depuis 2002

Je fais partie d'un réseau de correspondants TICE des grandes écoles de la région. Chaque année plusieurs rencontres sont planifiées afin de partager les expériences de chacun dans le domaine des TICE. Le réseau TICE fonctionne également au travers d'un espace collaboratif pour cette communauté servant d'espace de travail et de communication.

Membre du Conseil de Département ETN depuis 1990

Je participe depuis de nombreuses années au Conseil de Département auquel je suis affecté. Ce conseil assiste le directeur d'ETN dans la définition de la politique du département, aussi bien dans le contenu ou l'organisation de la pédagogie, que dans la vie générale du département (budget, demande de postes,...).

Membre du Conseil d'Administration de l'IMAR (Image, Mer, aménagement Régional) de 1992 à 1994

L'IMAR est une association qui contribue au développement de la géomatique dans l'Ouest. Elle organise un certain nombre d'actions, notamment le partage d'expériences et la diffusion des bonnes pratiques en ce qui concerne les Systèmes d'Information Géographiques.

Correspondant informatique pour le département ETN depuis 2000

J'assure l'interface entre le département ETN et le Service Informatique de l'école. Ma mission consiste à établir les besoins informatiques du département, en définissant les machines et les logiciels devant être déployés dans les salles informatiques. Nous nous concertons entre départements lors de réunions des correspondants.

Participation au recrutement des Élèves-ingénieurs de Polytech'Nantes depuis 1990

Je participe chaque année au recrutement des étudiants de l'école en évaluant les dossiers et en prenant part aux entretiens avec les candidats.

10. Enseignement

Depuis mon affectation au poste de Maître de Conférences en 1990 à Polytech'Nantes, j'effectue la grande part de mon activité d'enseignant dans le département *Electronique et Technologies Numériques* (département ETN de Polytech'Nantes, anciennement département SEII de l'IRESTE ; j'utiliserai par la suite l'actuelle dénomination). Il m'arrive d'intervenir dans le département Informatique de la même école. J'ai assuré de manière plus ponctuelle des cours dans différents établissements, tels que le CNAM ou l'Ecole des Mines de Nantes. J'interviens également dans le Master Professionnel de Bio-informatique de Nantes. J'enseigne enfin en formation continue dans le cadre du D.U. *Chef de Projet, Logiciel et Réseaux* à Nantes.

Le **nombre d'étudiants** dans une promotion de l'école est actuellement de **72**.

Je suis responsable de tous les cours que j'ai enseignés et j'ai conçu **8 supports de cours**.

La section suivante présente les disciplines enseignées dans l'ordre chronologique ainsi que des tableaux récapitulatifs des volumes horaires par année et par matière.

Disciplines enseignées

1990 – 1992	**Algorithmique et Programmation PASCAL** Département ETN, Polytech'Nantes Audience : Bac+3 Volume horaire annuel : TD (24 h), TP (36 h) *Notions abordées :* Bases de la programmation en PASCAL Complexité algorithmique Récursivité Structure de Données

1990 – 1993	**Algorithmique numérique** Département ETN, Polytech'Nantes Audience : Bac+3 Volume horaire annuel : Cours (2 h), TD (12 h), TP (16 h) *Notions abordées :* Résolution de systèmes linéaires Interpolation et intégration numérique Equations différentielles, racines d'une équation

1990 – 1999	**Programmation avancée ADA** Département ETN, Polytech'Nantes Audience : Bac+4 Volume horaire annuel : Cours (12), TP (32h) *Notions abordées :* Bases de la programmation en ADA Structure de Données Avancées Gestion des exceptions

1990 –	**Bases de données** Département ETN, Polytech'Nantes Audience : Bac+4 Volume horaire annuel : Cours (4,5 h),TP (32h) *Notions abordées :* Modélisation UML pour les bases de données Bases de données relationnelles Langage SQL

1990 – 1998	**Traitements d'images** Département ETN, Polytech'Nantes Audience : Bac+4 Volume horaire annuel : TP (96h) *Notions abordées :* Amélioration d'images Filtrage Morphologie mathématique

1992 – 1996	**Algorithmique et Programmation C++** Département ETN, Polytech'Nantes Audience : Bac+3 Volume horaire annuel : Cours (42 h), TD (24), TP (52 h) *Notions abordées :* Bases de la programmation en C++ Programmation Objet Complexité algorithmique Récursivité Structure de Données

1993 – 1996	**Architectures matérielles, modélisation VHDL** Volume horaire annuel : TP (8 h) *Notions abordées :* Modélisation d'un processeur Simulation en VHDL

1997 –	**Algorithmie et langage Java** Département ETN, Polytech'Nantes Audience : Bac+3 Volume horaire annuel : Cours (22 h), TD (22), TP (45 h) *Notions abordées :* Bases de la programmation en C++ Programmation Objet Complexité algorithmique Récursivité Structure de Données

1997 –	**Modélisation UML** Département ETN, Polytech'Nantes Audience : Bac+5 Volume horaire annuel : Cours(15h), TP (9h) *Notions abordées :* Amélioration d'images Filtrage Morphologie mathématique

2004 –	**Programmation Objet** Formation DUTIL, Polytech'Nantes Audience : Bac+4 Volume horaire annuel : Cours (18h), TP (12h) *Notions abordées :* Concepts de base Héritage et polymorphisme Traitement des exceptions

Récapitulatif

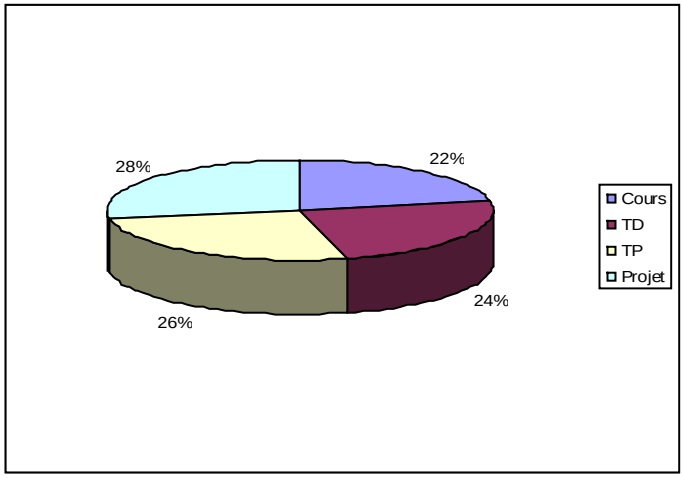
Le tableau ci-dessous présente les volumes horaires en équivalent TD sur les différentes années universitaires, regroupés par le niveau d'études (BAC +3 à BAC +5, ainsi que la formation continue (FC)). Les différences importantes de charge reflètent essentiellement l'évolution de la formation dans le département *Electronique et Technologies Numériques*. L'enseignement de la programmation a notamment considérablement évolué, tant dans le fond (j'ai activement milité pour l'introduction de la programmation objet dans les disciplines du département ainsi que l'utilisation du langage Java comme premier langage de programmation) que dans la forme (passage progressif de la 4ème vers la 3ème année de la programmation Objet). D'autres facteurs expliquent ces fluctuations : on peut citer la création de projets transversaux qui a nécessité la réduction du volume horaire de tous les enseignements magistraux, ainsi que la réforme de la semestrialisation.

La charge globale fluctue autour d'une moyenne de 256 heures eq. TD, la médiane étant à 243 heures. Ce volume important s'explique par la part non négligeable des charges non encadrées (projets techniques et projets transversaux), ce qui est le lot commun des écoles d'ingénieurs. Le pic de charge en 2002 est ainsi caractérisé par l'ajout de 102 h. eq. TD d'encadrement de projets divers soit environ un tiers de la charge globale. Par ailleurs, le déficit chronique en enseignant d'informatique dans le département entraîne une multiplication des séances, puisqu'un enseignant doit souvent encadrer de 2 à 4 groupes d'étudiants pour les mêmes travaux pratiques.

Année	**BAC +3**	**BAC +4**	**BAC +5**	**F. C.**	**Total**
1990 – 1991	68	113	39		220
1991 – 1992	76	137	75		288
1992 – 1993	120	108	102		330
1993 – 1994	95	129	38		262
1994 – 1995	96	92	42		230
1995 – 1996	106	86	29		221
1996 – 1997	106	118	45		269
1997 – 1998	102	68	60		230
1998 – 1999	123	70	50		243
1999 – 2000	Conversion thématique				
2000 – 2001	115	120	37		272
2001 – 2002	95	131	24		250
2002 – 2003	96	116	115		327
2003 – 2004	116	39	74	5	234
2004 – 2005	100	21	78	38	237
2005 – 2006	86	67	46	42	241
2006 – 2007	86	90	42	54	272
2007 – 2008	86	71	42	51	250
2008 – 2009	86	59	42	44	231
TOTAL	*1758*	*1635*	*980*	*234*	*4607*

Tableau récapitulatif des charges d’enseignements

La répartition entre cours, TDs, Tps et Projets est relativement homogène et stable d'une année à l'autre. Sur la figure ci-dessous, on observe que ce sont les projets qui constituent une part prépondérante dans le volume total de la charge (28 %).



Pourcentage des différents types d'enseignement.
Les données représentent une moyenne évaluée sur les cinq dernières années d'enseignement. Ce calcul est basé sur les heures effectives enseignées.