HAL
open science

Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs

Jérôme Azé

To cite this version:

Jérôme Azé. Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs. Apprentissage [cs.LG]. Université Paris Sud - Paris XI, 2012. tel-00763947

HAL Id: tel-00763947

https://theses.hal.science/tel-00763947v1

Submitted on 17 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs

Habilitation à Diriger des Recherches

(Spécialité Informatique)

Université Paris-Sud

présentée et soutenue publiquement le 16 novembre 2012

par

Jérôme Azé

Rapporteurs :	Olivier Lichtarge	Professeur, Baylor College of Medecine, Houston, Texas, USA
	Stan Matwin	Professeur, Université d'Ottawa, Canada
	Amedeo Napoli	Directeur de Recherche CNRS, LORIA, Nancy, France
Examinateurs :	Michel Beaudouin-Lafon	Professeur, Université Paris Sud, Orsay, France
	Antoine Cornuéjols	Professeur, Agro Paris Tech, Paris, France
	Christine Froidevaux	Professeur, Université Paris Sud, Orsay, France
	Engelbert Mephu Nguifo	Professeur, Université Blaise Pascal, Clermont-Ferrand, France
Marraine :	Anne Poupon	Directrice de Recherche CNRS, INRA Tours, France

Laboratoire de Recherche en Informatique – CNRS UMR 8623
Équipe associée INRIA Saclay – Île-de-France – Université Paris-Sud,
bât. 650, F-91405, Orsay Cedex France
Secrétariat : Tél.: +33 (0)1 69 15 66 29 – Fax : +33 (0)1 69 15 65 79 70
http://www.lri.fr

Mis en page avec la classe thloria.

TABLE DES MATIÈRES

Introduction **1**

1 Apprentissage artificiel et fouille de données **5**

A Contexte général 6

B Apprentissage supervisé 6

1 Paradigme de l'apprentissage supervisé 6

2 Critères d'évaluation 7

2.1 Critères d'évaluation globaux 7

2.2 Critères d'évaluation "locaux" 8

3 ROGER où comment apprendre à maximiser l'aire sous la courbe ROC par algorithme évolutionnaire 10

4 Ensemble learning 11

4.1 Stacking 12

4.2 Boosting 12

4.3 Bagging 12

C Apprentissage non-supervisé 16

1 Clustering 16

2 Extraction de motifs fréquents et règles d'association 16

3 Règles d'association et mesures de qualité 17

2 Prédiction d'interactions Protéine-Protéine **19**

A Contexte général 20

1 Principe général de l'approche 23

2 Apprentissage du modèle 23
B Résultats 25
1 Prédiction des paires de protéines en interaction 25
2 Élagage des paire de protéines 27
C Conclusion 27

3 Amarrage Protéine-Protéine **31**
A Introduction 32
B Modélisation atomique *vs* modélisation gros-grain 32
1 Modélisation gros-grain par diagramme de Voronoï et apprentissage de fonctions de score 33
2 Génération des conformations à l'aide des diagrammes de Voronoï 34
3 Génération : Hex, évaluation : Voronoï + apprentissage 34
C Compétition CAPRI 35
1 Critères d'évaluation de CAPRI 35
2 Résultats 36
D Conclusion 36

4 Conclusion, perspectives **39**
A Ré-annotation fonctionnelle de génomes et annotation de mutants 39
B Cross-docking 40
C Réseau d'interaction protéine-protéine 41
D Interactions protéine-ARN 41

Bibliographie **43**

Partie I Curriculum vitae **53**

Notice Individuelle **55**
Situation Professionnelle 55
Informations Personnelles 55
Domaine de recherche 55

Parcours et Formation Universitaire **57**
Situation actuelle 57
Titres Universitaires 57

Parcours post-doctoral 58
Thèse 58
Parcours de premier et second cycle 58

Activités d'enseignements **61**
Principaux Enseignements 61
Production de documents pédagogiques 61

Activités liées à l'administration **63**
Responsabilités pédagogiques 63
Commissions de spécialistes 63
Autres responsabilités – vie du laboratoire 63

Animations scientifiques majeures **65**
Jury de Thèse 65
Comité de Programme de Conférences internationales 65

Challenges internationaux de fouilles de données **67**
Participation 67
Création 67

Encadrement de Post-Docs, de Thèses et de Masters Recherche **69**
Post-Doc 69
Thèse 69
Encadrement de Stage de Master Recherche 70

Collaborations **71**
Industrielles 71
Académiques 71

Publications Scientifiques **73**
Revues internationales avec comité de lecture 73
Revues nationales avec comité de lecture 74
Chapitres de livres 74
Éditions d'ouvrages 74
Actes de conférences internationales avec comité de lecture 75
Workshops Internationaux 75
Actes de Challenges Internationaux 76

Actes de conférences nationales avec comité de lecture 76
Workshops Nationaux 77
Posters 77

INTRODUCTION

J'ai débuté ma carrière de chercheur en septembre 2000 lorsque j'ai été recruté en thèse par Yves Kodratoff au LRI dans l'équipe Inférence & Apprentissage qu'il dirigeait alors. Mon sujet de thèse concernait initialement la fouille de données et plus particulièrement l'étude des mesures de qualité des règles d'association afin de mieux définir la notion d'"intérêt" tel qu'un expert pourrait se la figurer. L'un des domaines d'application qui intéressait Y. Kodratoff était la fouille de textes et j'ai pu naturellement évaluer l'impact des mesures d'extraction que j'étudiais en les appliquant sur des problématiques textuelles ("text mining"). J'ai notamment collaboré avec Mathieu Roche qui avait débuté sa thèse en septembre 2001 également sous la direction d'Y. Kodratoff. Cette collaboration dura plusieurs années et fut fructueuse [KARMT03, RAKS04, AAH+04, ARS05a, ARKS05]. Mathieu et moi avons notamment créé en 2005 le DÉfi Fouille de Textes (DEFT[1]) [AAA+05]. Il s'agit d'un défi dédié aux textes rédigés en langues française. La 8ème édition de DEFT s'est déroulée en 2012 dans le cadre de la conférence TALN.

Ensuite, j'ai commencé à m'intéresser aux problématiques de fouille de données appliquées à des domaines liés à la biologie et la médecine à la fin de ma thèse. À cette période, j'ai eu l'occasion de travailler avec Noël Lucas qui préparait sa thèse de médecine en réalisant en parallèle un M2R d'informatique au LRI (Université Paris-Sud) dont le stage était encadré par Michèle Sebag. L'un des objectifs de sa thèse était d'évaluer l'impact des approches de fouille de données dans le domaine de la médecine et plus particulièrement sur l'usage des algorithmes évolutionnaires pour apprendre des modèles permettant de prédire le risque cardio-vasculaire. Dans le cadre de cette collaboration, N. Lucas incarnait l'expert des données médicales, et moi l'expert en fouille de données. Cette collaboration fut fructueuse et donna lieu à plusieurs publications, notamment dans le cadre de challenges de fouille de données dédiées à des problématiques bio-médicales [LAS02, SAL03, ALS03, ALS04].

Pendant cette même période, j'ai eu l'occasion de rencontrer Julie Bernauer dans un jury de master. Elle était en début de thèse, j'avais soutenu récemment la mienne et nous avons donc discuté de nos sujets respectifs. Nous nous sommes rapidement aperçu que nous pourrions très certainement collaborer. Elle travaillait sur la prédiction d'interaction physique entre protéines (l'amarrage protéine-protéine ou docking) et l'approche que nous avions développé avec N. Lucas pourrait certainement être adaptée pour la problématique de l'amarrage protéine-protéine. Ce furent là mes premiers pas en bioinformatique.

Depuis 2005, date de mon recrutement sur un poste de maître de conférences dans l'équipe Bioinformatique du LRI, j'ai consacré la quasi totalité de mes travaux de recherches à l'étude des interactions

[1] `http ://deft.limsi.fr/`

entre protéines.

Les protéines sont des macromolécules biologiques dont le rôle est primordial dans tous les processus du vivant, de la machinerie cellulaire au maintien des parois cellulaires, ou à la structure des tissus osseux. La fonction d'une protéine est entièrement dépendante de sa structure tridimensionnelle, qui peut être décrite à quatre niveaux de complexité et d'organisation (voir figure 1) :

- la structure primaire : Une protéine est un assemblage linéaire d'acides aminés, appelé séquence primaire, ou chaîne polypeptidique.
- la structure secondaire : La chaîne polypeptidique est capable de former des arrangements locaux périodiques, appelés structures secondaires.
- la structure tertiaire : La structure tertiaire décrit le repliement dans l'espace des éléments de structure secondaire d'une chaîne polypeptidique unique. Dans les conditions physiologiques, ce repliement est unique, spontané ou aidé dans sa maturation par d'autres protéines : les chaperonnes.
- et la structure quaternaire : Alors que certaines protéines sont constituées d'une chaîne polypeptidique unique, d'autres sont fonctionnelles uniquement sous forme d'oligomères, c'est-à-dire l'assemblage de plusieurs chaînes polypeptidiques.

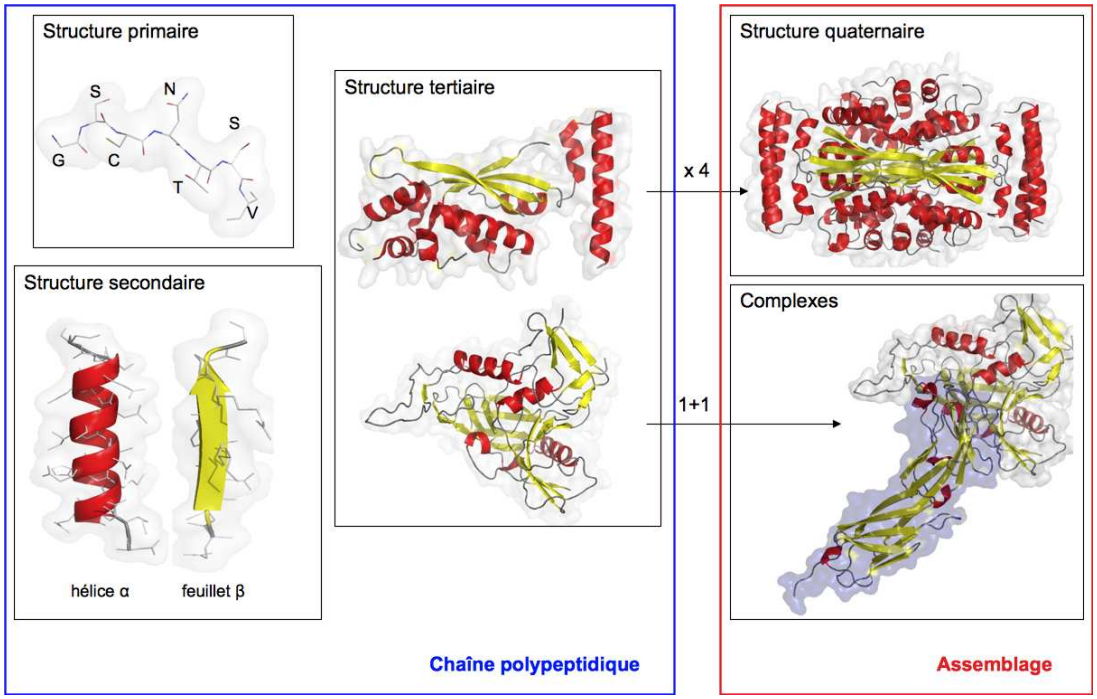

FIGURE 1 : Les quatre structures d'une protéine.

La structure tridimensionnelle d'une protéine, y compris ses interactions avec différents partenaires tels que des petites molécules ou d'autres macromolécules, peut être déterminée expérimentalement par cristallographie aux rayons X ou par Résonance Magnétique Nucléaire (RMN). Cependant, même si ces techniques ont connu des avancées fondamentales, notamment grâce aux projets de génomique structurale, cette détermination reste délicate voire impossible. De plus, le nombre considérable de protéines connues, et la combinatoire découlant des interactions entre elles, et avec d'autres molécules, rendent inenvisageable le recours systématique à l'expérimentation (environ 4100 protéines pour *Bacillus subtilis* et 26000 pour l'homme). L'approche *in silico*, qui consiste à **prédire**, et non plus à déterminer, la structure tridimensionnelle des protéines et des complexes, peut permettre de répondre à cette problé-

matique. La modélisation *ab initio* permet aujourd'hui de prédire, avec un succès de plus en plus grand, le repliement d'une protéine isolée à partir de sa séquence [MFKT11]. D'autre part, les efforts d'une communauté grandissante de chercheurs, communauté dans laquelle nous nous plaçons, portent sur la prédiction de la conformation des complexes macromoléculaires à partir des structures des partenaires isolés (structures déterminées expérimentalement ou prédites) [LG08, RK00, DBB03, BBAP11].

La plupart des protéines réalisent leur fonction biologique via l'interaction avec un autre partenaire (protéine, ARN, *etc*). Lors de cette interaction, la structure physique de la protéine peut être modifiée de façon significative. Trois types de déformations sont actuellement admis par la communauté : (i) key-and-lock où la protéine et son interactant se comportent comme un mécanisme clé-serrure avec une parfaite adéquation structurale de l'un envers l'autre, (ii) "induced fit" où la protéine se déforme physiquement lors de l'interaction pour adopter la conformation propice à l'interaction avec son partenaire, et (iii) "selected fit" où la protéine adopte naturellement plusieurs conformations différentes et l'interactant sélectionne la conformation adaptée [WvD09].

La connaissance des structures en trois dimensions des protéines est donc un véritable challenge si l'on souhaite comprendre finement les mécanismes fonctionnels associés aux protéines, pouvoir les prédire et pouvoir les inhiber ou les activer en construisant des interactants spécifiques d'une protéine donnée et d'une fonction biologique donnée (drug design).

Depuis le début des années 2000, de nombreux projets de biologie structurale ont vu le jour (PSI, 1998 ; MCSG, 2000 ; SGC, 2004 ; ...). Ces projets ont pour ambition soit d'arriver à collecter l'ensemble des structures et des interactions entre protéines d'un génome donné comme par exemple le projet BSCG[2] qui se focalise sur l'étude exhaustive de deux génomes pathogènes de l'homme et de l'animal : *Mycoplasma genitalium* et *Mycoplasma pneumoniae*. Soit d'améliorer globalement les processus de détermination de la structure des protéines et des interactions entre protéines, comme par exemple le projet PSI[3] dont l'objectif clairement affiché est de réduire les coûts et le temps nécessaire pour obtenir la structure 3D d'une protéine. L'un des objectifs à long terme de PSI est de résoudre 10 000 structures de protéines en 10 ans et de faire en sorte que la structure des protéines au niveau atomique puisse être prédite directement à partir des connaissances accumulées ou directement à partir de la séquence d'ADN associée à la protéine. Actuellement, plus de 20 millions de protéines sont recensées[4] et environ 80000 structures sont résolues[5]. L'amélioration des techniques de séquençage (NGS, Next Generation Sequencing) est telle qu'avec l'arrivée prochaine des NGS de 3ème génération[6], nous allons crouler sous une masse de génomes entièrement séquencés sans pour autant être capable d'obtenir la structure physique des protéines associées.

De nombreux travaux récents montrent que la prédiction de l'interactome complet (l'ensemble des interactions entre protéines) de tout génome nouvellement séquencé représente l'un des défis majeurs de la bioinformatique [ML12, RHJ+12, VCB11].

Nous nous inscrivons donc nous aussi dans la perspective d'améliorer la qualité de la prédiction des structures des protéines, notamment en prédisant de manière fiable les interactions entre protéines. Les travaux auxquels j'ai participé depuis bientôt 10 ans sont dédiés à deux aspects des interactions protéine-protéine :

- prédiction des interactions physiques entre protéine pour permettre de prédire l'agencement structurale de deux protéines en interaction (voir le chapitre 3 consacré à l'amarrage protéine-protéine)

[2] http ://www.strgen.org/
[3] http ://www.nigms.nih.gov/Research/FeaturedPrograms/PSI/
[4] http ://www.ebi.ac.uk/uniprot/TrEMBLstats/
[5] http ://www.pdb.org/pdb/statistics/holdings.do
[6] http ://the-scientist.com/2012/06/01/sons-of-next-gen/

– prédiction de l'existence d'une interaction entre deux protéines directement à partir de leur séquences d'ADN (voir le chapitre 2 consacré à la prédiction des interactions protéine-protéine)

Nous présenterons dans la suite de ce document les méthodes proposées et mises en œuvre pour apporter des réponses à ces deux problématiques.

Nous distinguerons deux familles d'approches :

– les approches reposant sur une modélisation des lois physico-chimiques régissant les interactions entre protéines et conduisant souvent à des approches minimisant une fonction énergétique ;
– et les approches reposant sur des mécanismes d'apprentissage supervisé de modèles permettant de prédire au mieux l'interaction entre deux protéines.

La première famille cherche à modéliser finement les interactions physico-chimiques impliquées dans l'interaction physique de deux protéines. Les différentes modélisations obtenues correspondent souvent à des combinaisons de fonctions énergétiques dont les paramètres ont été ajustés sur la bases des complexes connus. Plusieurs logiciels intègrent de telles approches [LG08, DBB03].

Dans le cadre de nos travaux, nous nous situons dans la deuxième famille d'approches, à savoir l'apprentissage de modèles permettant de prédire au mieux l'interaction entre deux protéines, soit directement à partir de leurs séquences d'acides aminés (structure primaire), soit à partir de leurs structures en 3 dimensions.

Bien que notre objectif soit toujours de prédire les interactions entre protéines, la nature de la prédiction diffère en fonction de la nature de la description des protéines. Ainsi, à partir de deux séquences d'acides aminés, nous pourrons prédire si deux protéines sont susceptibles d'interagir et nous pouvons quantifier cette propension à interagir. Le chapitre 2 est consacré à cette phase de la prédiction des interactions protéine-protéine.

Ensuite, étant donné un couple de protéines dont la probabilité d'interaction est élevée, nous pouvons prédire, à partir des structures 3D de chaque protéine, la nature exacte de l'interaction physique des deux protéines. Le chapitre 3 présente nos travaux dans ce domaine en les positionnant par rapport à ceux de la communauté.

Bien que la nature des données étudiées diffèrent, l'apprentissage des modèles prédictifs repose sur des méthodes similaires. Le chapitre 1 est consacré à la présentation détaillée des approches que nous avons étudiées et mises en œuvre dans le cadre de la prédiction d'interactions protéine-protéine.

Enfin, le chapitre 4 présente des perspectives à ces travaux de recherche, dont les résultats sont tels qu'il devient réellement possible d'envisager des applications de “drug design”.

Sommaire

A Contexte général
B Apprentissage supervisé
1 Paradigme de l'apprentissage supervisé
2 Critères d'évaluation
3 ROGER où comment apprendre à maximiser l'aire sous la courbe ROC par algorithme évolutionnaire
4 Ensemble learning
C Apprentissage non-supervisé
1 Clustering
2 Extraction de motifs fréquents et règles d'association
3 Règles d'association et mesures de qualité

CHAPITRE

1

APPRENTISSAGE ARTIFICIEL ET FOUILLE DE DONNÉES

Principales publications du chapitre

[SLA03] **ROC-based Evolutionary Learning : Application to Medical Data Mining**, EA 2003
[ALS04] **A Genetic ROC-based Classifier**, Challenge ICML 2004
[ARKS05] **Preference Learning in Terminology Extraction. A ROC-based approach**, ASMDA 2005
[Azé03] **Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances**, EGC 2003
[ALLV07] **A study of the robustness of association rules**, DMIN 2007
[LA07] **A New Probabilistic Measure of Interestingness for Association Rules Based on the Likelihood of the Link**, Quality Measures in Data Mining 2007

A Contexte général

L'ouvrage "Apprentissage artificiel – Concepts et algorithmes" d'Antoine Cornuéjols et de Laurent Miclet édité en 2002, puis réédité en 2010 [CM02] dresse un panorama relativement exhaustif de l'état de l'art en apprentissage.

Il propose une répartition des applications relevant du domaine de l'apprentissage artificiel selon deux axes : (i) reconnaissance des formes et (ii) extraction de connaissances à partir des données. Il est également possible de partitionner l'apprentissage automatique en fonction de la nature des données étudiées : **apprentissage supervisé** où les données sont partiellement labellisées (étiquetées) *vs* **apprentissage non supervisé** (données sans labels).

Dans le cadre de l'analyse des protéines (prédiction d'interactions protéine-protéine et amarrage protéine-protéine), nous nous sommes essentiellement focalisés sur des méthodologies relevant du domaine de l'apprentissage supervisé : apprentissage d'un modèle prédictif à partir des données connues. Mais nous avons également utilisé et exploité des spécificités du domaine de l'apprentissage non supervisé (les **mesures de qualité**) classiquement utilisées pour évaluer l'intérêt des règles d'association afin d'enrichir la description des données manipulées. En effet, certaines données contiennent nativement trop peu de descripteurs et l'enrichissement des données par des descripteurs obtenus à partir des données initiales peut mettre en évidence certaines informations cachées dans les données. C'est le cas par exemple en médecine où un patient peut être décrit par sa taille et son poids. Mais souvent l'utilisation de l'IMC (Indice de Masse Corporelle) qui est égal à $\frac{poids}{taille^2}$ permet d'obtenir un indicateur de sur-poids immédiatement interprétable par les médecins. La définition de nouveaux descripteurs d'une telle nature suppose d'avoir une connaissance avancée du domaine étudié. N'ayant pas forcément une telle connaissance, et ne souhaitant pas injecter trop d'*a priori* dans les données, nous avons proposé une méthode d'enrichissement des données entièrement automatique et dédiée aux données étudiées (ici des paires de protéines, voir chapitre 2).

La suite de ce chapitre est organisé selon ces deux axes : apprentissage supervisé (voir section B et apprentissage non supervisé (voir section C).

B Apprentissage supervisé

1 Paradigme de l'apprentissage supervisé

Le paradigme de l'apprentissage supervisé peut se résumer de la manière suivante : *étant donné un ensemble d'exemples étiquetés, apprendre un modèle capable de prédire au mieux les étiquettes de nouveaux exemples*.

Soient $\mathcal{X}$ l'ensemble des **exemples** (ou données) et $\mathcal{Y}$ l'ensemble des **étiquettes** (notées aussi **classes**) pouvant être associées aux exemples. Dans le cadre des travaux présentés dans ce manuscrit, seules des étiquettes binaires ont été considérées : $\mathcal{Y} = \{-1, +1\}$ (notées également $\{+, -\}$).

Les données peuvent se répartir en deux catégories :
- les données déjà étiquetées. Ces données sont en général présentes en faible quantité car il est souvent très coûteux d'obtenir l'étiquette associée à une donnée (par ex : obtenir la structure d'un complexe protéine-protéine par une expérience de crystallographie). Ces données seront utilisées pour apprendre un modèle permettant de prédire les étiquettes de nouveaux exemples.
- les données non étiquetées. Il est en général aisé d'obtenir de telles données.

L'ensemble des données déjà étiquetées est nommé **l'ensemble d'apprentissage**. Nous le désignerons par $\mathcal{A} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ avec $\boldsymbol{x}_i \in \mathbb{R}$ et $y_i \in \mathcal{Y}, \forall i \in \{1, \ldots, n\}$. $\boldsymbol{x}$ est un vecteur de dimension d où chaque dimension représente l'une des caractéristiques de l'exemple $\boldsymbol{x}$.

Nous verrons dans les chapitres suivants que la nature des exemples peut varier :
- Paires de protéines décrites par des informations phylogénétiques (voir chapitre 2)
- Paires de protéines décrites par des informations issues d'une représentation en 3 dimensions des protéines (voir chapitre 3)

L'**apprentissage** se déroule classiquement en trois phases :
- Apprentissage d'un modèle permettant de prédire au mieux les données d'apprentissage ;
- Évaluation de ce modèle sur un jeu de données extrait du jeu d'apprentissage (en utilisant par exemple une procédure de validation-croisée ou de "leave-one-out") ;
- Test du modèle obtenu sur un jeu de données disjoint du jeu d'apprentissage.

Le processus d'évaluation des performances d'un modèle appris nécessite d'utiliser des données non utilisées pour l'apprentissage afin de ne pas biaiser les évaluations de performances. Deux processus sont classiquement utilisés : la **validation-croisée** et le **Leave-One-Out**.

L'évaluation par **validation-croisée** consiste à partitionner les données d'apprentissage $\mathcal{A}$ en k parties disjointes, d'apprendre sur l'union de $k-1$ parties et d'évaluer les performances sur la partie non utilisée. Ce processus est itéré k fois, ainsi tous les exemples de $\mathcal{A}$ auront été utilisés une fois en test et $k-1$ fois en apprentissage. Le choix de la valeur de k dépend de la taille des données. Les valeurs classiquement utilisées sont $k = 3$ ou $k = 10$.

L'évaluation par **Leave-One-Out** est une généralisation de la validation-croisée avec $k = n$. Ainsi, pour chaque exemple, un modèle est appris à partir de l'intégralité des données sauf l'exemple de test. Ce protocole d'évaluation est utilisé lorsque les données sont peu nombreuses et que le recours à la validation-croisée conduirait à se priver d'une trop grande partie des données pour l'apprentissage. Dès que les données sont trop volumineuses, le recours à cette méthode n'est plus viable car le coût de calcul devient rapidement prohibitif (apprentissage de n modèles).

Les critères d'évaluation utilisés pour mesurer les performances des modèles appris sont essentiels dans tout processus d'apprentissage. De nombreux critères d'évaluation ont été proposés dans la littérature et nous présentons ci-après les critères les plus fréquemment utilisés et notamment ceux que nous manipulerons dans la suite de ce document.

2 Critères d'évaluation

Tout d'abord, lorsqu'un modèle prédictif est appliqué sur un jeu de données, nous pouvons mesurer, pour chaque étiquette, le nombre d'exemples correctement associés à cette étiquette, ainsi que le nombre d'exemples qui y sont incorrectement associés. Ces informations sont rassemblées dans une matrice nommée la **matrice de confusion**.

Dans le cadre de la classification binaire, la matrice de confusion peut se représenter sous la forme suivante :

2.1 Critères d'évaluation globaux

À partir de cette matrice de confusion, de nombreux critères d'évaluation peuvent être calculés. Parmi les plus utilisés, nous pouvons citer :

		Réel	
		+	-
Prédit	+	VP	FP
	-	FN	VN

TABLE 1.1 : où VP représente les Vrais Positifs, FP les Faux Positifs, FN les Faux Négatifs et VN les Vrais Négatifs. Cette matrice de confusion restreinte à un problème à deux classes peut être étendue à un problème à n classes. La notion de Faux Positifs ou Faux Négatifs doit alors également être étendue.

- La **précision** $P = \frac{VP}{VP+FP}$. Cette mesure représente le pourcentage de prédictions correctes associées à la classe positive (la même mesure peut être définie pour la classe négative) ;
- Le **rappel** $R = \frac{VP}{VP+FN}$. Cette mesure représente le pourcentage d'exemples positifs correctement prédits positifs (de la même manière que pour la précision, il est possible de définir cette métrique pour la classe négative) ;
- Le $\boldsymbol{F}_{score}(\beta) = \frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$ qui permet d'agréger en une seule métrique la précision et le rappel ; Le paramètre β permet de pondérer la précision *vs* le rappel. Si $\beta < 1$, le poids de la précision devient plus important, inversement, lorque $\beta > 1$, le poids de la précision diminue. Lorsque $\beta = 1$ la précision et le rappel ont la même importance. La valeur de β est très fréquemment fixée à 1 ;
- L'**accuracy** $Acc = \frac{VP+VN}{VP+VN+FP+FN}$ qui permet d'évaluer la performance "globale" d'un classifieur. Cette mesure représente le pourcentage de prédictions correctes toutes classes confondues ;
- La **sensibilité** $Se = \frac{VP}{VP+FN}$ qui est égale au rappel de la classe positive. Cette mesure est issue du domaine du traitement du signal et est largement utilisée dans le domaine médical ;
- La **spécificité** $Sp = \frac{VN}{FP+VN}$ qui correspond au rappel des négatifs. Cette mesure est également issue du domaine du traitement du signal. Son utilisation dans le domaine médical est toujours associée à la sensibilité. Ces deux mesures permettent d'évaluer l'efficacité d'un nouveau test médical en indiquant sa capacité à effectuer à la fois des prédictions correctes pour les positifs (sensibilité), tout en couvrant peu de négatifs (capacité évaluée par $1 - Sp$).

Toutes ces métriques fournissent une vision "globale" des performances d'un classifieur car elles résument en une unique valeur le comportement du classifieur sur l'ensemble des données.

D'autres métriques ou critères d'évaluation ont été proposés pour permettre d'obtenir une vision plus fine des performances d'un ou plusieurs classifieurs.

2.2 Critères d'évaluation "locaux"

Il est notamment devenu évident, depuis les années 2000, qu'il était insuffisant d'évaluer les performances d'un classifieur uniquement avec la précision et le rappel. De nouvelles métriques se sont rapidement imposées dans la communauté [LHZ03b, FF03]

Ces métriques peuvent être réparties en deux catégories :
- celles qui permettent de comparer plusieurs classifieurs entre eux (ou plusieurs réglages d'un même classifieur) ;
- celles qui permettent d'évaluer les classifieurs associant un score à chacune de leur prédiction. Ce score, qui peut être assimilé à un degré de confiance dans la prédiction effectuée, permet alors d'ordonner les prédictions et ainsi d'obtenir plus d'informations qu'une "simple" étiquette.

La visualisation du compromis entre précision et rappel fait ainsi partie de la première catégorie (voir figure 1.1). Ce type de courbe permet de comparer très rapidement les performances de plusieurs classi-

fieurs. Nous pouvons ainsi constater que les classifieurs C_1 dominent tous les autres classifieurs. Même si en l'absence d'*a priori* sur le compromis précision/rappel recherché, nous ne pouvons pas choisir un classifieur parmi l'ensemble des classifieurs de type C_1, il est évident qu'il est préférable d'utiliser C_1 plutôt que C_2 ou C_3. Par contre, d'autres critères peuvent entrer en considération et conduire l'utilisateur à privilégier les classifieurs de type C_2 : temps nécessaire pour prédire, coût financier (logiciels libres *vs* logiciels payants), . . .

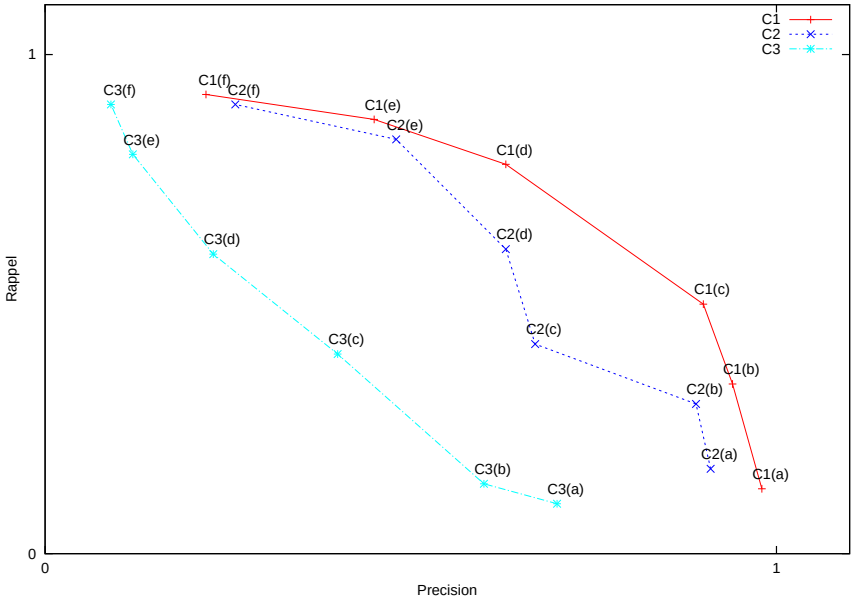

FIGURE 1.1 : Visualisation du compromis précision/rappel pour trois classifieurs différents (C1, C2 et C3), ainsi que pour six paramétrages différents de chaque classifieurs (a, b, c, d, e et f). Cette courbe présente les trois fronts de Pareto. Chaque point représente les performances d'un classifieur pour un paramétrage donné.

L'analyse de ces courbes est formalisée par la notion de **front de Pareto** et de **Pareto optimalité** [Buc62, AS01, Deb03].

Le front de Pareto est défini par l'ensemble des approches qui sont telles qu'aucune autre approche ne présente de meilleurs résultats pour tous les critères étudié (par exemple précision et rappel). Les approches qui ne sont pas sur le front de Pareto sont dites **dominées**.

La seconde catégorie de métriques concerne les classifieurs ayant la capacité d'associer à chaque prédiction une quantité numérique reflétant la confiance dans la prédiction, ou directement l'estimation de la probabilité d'appartenance à une classe (en général la classe positive pour un classifieur binaire).

La **courbe ROC** (Receiver Operating Characteristic) [VC06, Met78] permet de visualiser le compromis entre la sensibilité et la spécificité. La courbe ROC associée à un classifieur idéal est constituée de deux segments : un premier segment reliant le point (0,0) au point (0,1) correspondant aux exemples positifs parfaitement ordonnés puis un second segment reliant le point (0,1) au point (1,1) correspondant aux exemples négatifs ayant tous des scores inférieurs aux scores des exemples positifs. Cette courbe représente un classifieur ayant la capacité de séparer parfaitement les positifs des négatifs (voir figure 1.2).

Les courbes ROC permettent de visualiser rapidement les performances d'un ou plusieurs classifieurs. Afin de pouvoir comparer des classifieurs, notamment dans un cadre de recherche du meilleur

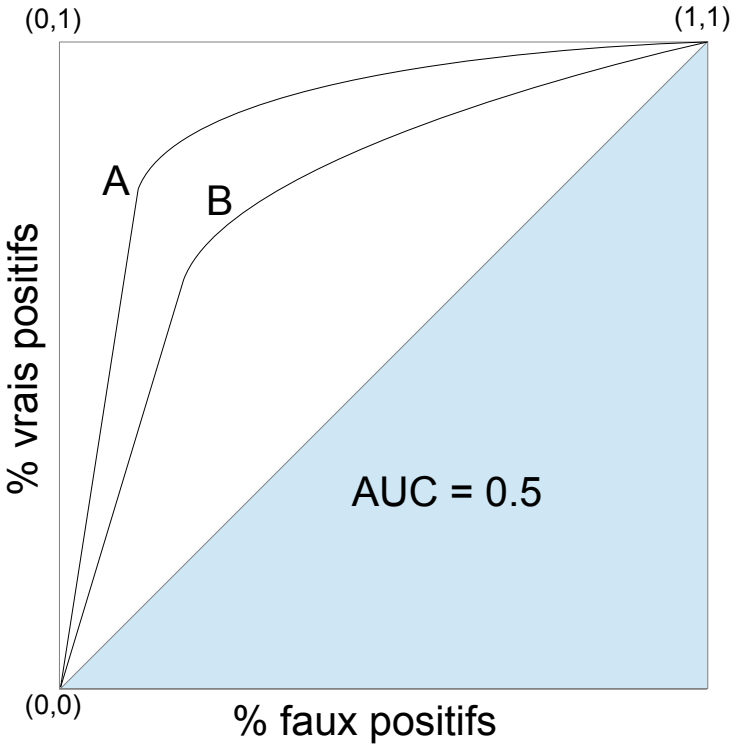

FIGURE 1.2 : Les courbes ROC associées aux classifieurs A et B de visualiser la supériorité du classifieur A par rapport au classifieur B.

classifieur (selon un ou plusieurs critères d'évaluation), il est très utile de pouvoir comparer numériquement les performances de ces classifieurs.

L'aire sous la courbe ROC (**AUC**) est très largement utilisée pour comparer les performances de plusieurs classifieurs. [LHZ03b, LHZ03a] ont montré que l'aire sous la courbe ROC (AUC, "Area Under the Curve") est une métrique plus fiable que l'accuracy pour comparer deux classifieurs.

De nombreux classifieurs "classiques" ont été adaptés pour pouvoir intégrer l'optimisation de l'AUC dans leur critère d'apprentissage comme les SVM [Rak04] ou les arbres de décision [FFHO02].

Nous présentons dans la section suivante, notre contribution à cet effort : un algorithme évolutionnaire apprenant à maximiser l'aire sous la courbe ROC [ALS04].

3 ROGER où comment apprendre à maximiser l'aire sous la courbe ROC par algorithme évolutionnaire

Dans le cadre d'une collaboration avec Michèle Sebag et Noël Lucas (médecin), j'ai eu l'occasion de participer à la conception et au développement d'un algorithme évolutionnaire dont l'objectif est d'apprendre des fonctions d'ordonnancement d'exemples sous contrainte d'optimiser l'aire sous la courbe ROC associée à l'ordre induit par la fonction apprise.

En reprenant les notations précédemment introduites, l'algorithme ROGER (ROc based Genetic learnER) apprend des fonctions de la forme : $f(\boldsymbol{x}_i) = \sum_j w_j \times \boldsymbol{x}_i(j)$ où $\boldsymbol{x}_i(j)$ représente la valeur de la $j^{\text{ème}}$ composante de l'exemple $\boldsymbol{x}_i$. L'algorithme apprend les poids w_j tels que $\sum_i rang_f(\boldsymbol{x}_i) \times \mathbb{1}_{y_i=+1}$ soit minimale (où $rang_f(\boldsymbol{x}_i)$ correspond au rang de l'exemple $\boldsymbol{x}_i$ induit par la fonction f, et $\mathbb{1}_{y_i=+1}$

correspondant à la fonction indicatrice qui vaut 1 si la classe de x_i est positive et 0 sinon).

Il est aisé de montrer qu'une fonction maximisant l'aire sous la courbe ROC minimise la somme des rangs des exemples positifs qu'elle ordonne.

Cet algorithme a été appliqué avec succès à divers domaines : la prédiction du risque cardio-vasculaire [SAL03, SLA03, ALS03], l'extraction de termes textuels [ARKS05, ARKS04] et la prédiction d'interaction physique entre protéines [BAJP07, BAJP05, BPAJ05].

Comme tout algorithme évolutionnaire, ROGER débute son apprentissage à partir d'un ensemble de poids choisis aléatoirement. Cette initialisation peut induire un biais dans les performances et il convient donc d'effectuer plusieurs apprentissages en variant les tirages aléatoires initiaux afin d'estimer au mieux les performances de l'approche. Sachant que ROGER, comme la plupart des algorithmes évolutionnaires, ne fournit que la meilleure fonction apprise à l'issue de l'évolution, et que nous devons effectuer de multiples apprentissages pour obtenir une évaluation fiable des performances, nous avons proposé une variante de ROGER qui permet d'agréger les différences fonctions apprises. Nous nous situons ainsi dans le cadre du **Bagging** [Bre96] et plus généralement de l'**Ensemble Learning** [Qui96, OM99, BK98], approches détaillées dans la section suivante.

Contraitement à [GSST07] qui exploite l'ensemble des solutions parcourues par un algorithme évolutionnaire pour mettre en place une stratégie d'ensemble learning, nous avons proposé d'agréger *a posteriori* les fonctions apprises par ROGER selon le principe suivant :

1. apprentissage d'un ensemble de fonctions par ROGER
2. application de chaque fonction sur l'ensemble des exemples à prédire : obtention d'un ensemble de rangs pour chaque exemple
3. le rang final de chaque exemple correspond au rang médian de l'exemple considéré par rapport à l'ensemble des rangs qu'il a obtenu avec les diverses fonctions

L'un des avantages de cette fonction d'agrégation est qu'il est possible de retrouver la fonction ayant fourni le score (et donc le rang) d'un exemple particulier. Cette fonctionnalité peut s'avérer particulièrement utile dans un cadre applicatif où l'analyse des résultats est associée à une compréhension fine du classifieur.

Cette approche a été testée avec succès dans le cadre de l'extraction de la terminologie [ARS05b].

Dans un contexte plus général, nous présentons dans la section suivante le paradigme de l'ensemble learning et les trois de ses principales mise en œuvre : le "stacking", le "boosting" et le "bagging" .

4 Ensemble learning

Le paradigme de l'ensemble learning consiste à apprendre plusieurs classifieurs dans le but de résoudre le même problème. Par opposition à l'apprentissage "classique" où un seul classifieur est appris, l'ensemble learning apprend un ensemble de classifieurs et les combine pour obtenir un **méta-classifieur**.

Les premiers travaux consacrés à l'ensemble learning datent des années 1990 [Han90, Wol92, KV95] De mon point de vue, trois paramètres ont contribué à démocratiser l'ensemble learning depuis 1990.

1. la démocratisation de la puissance de calcul rendant accessibles l'apprentissage et l'utilisation de nombreux modèles pour résoudre le même problème ;
2. l'émergence de masses de données à traiter qui s'avèrent souvent non apprenables par un unique classifieur ;

3. un certain mimétisme du comportement des experts humains où dans de nombreux domaines (médecine, aérospatiale, nucléaire, *etc*) la prise de décision résulte souvent d'un consensus entre différents experts.

Diverses formes d'ensemble learning peuvent être citées. Elles divergent principalement par la fonction d'aggrégation mise en œuvre pour obtenir le méta-classifieur. Ces divergences peuvent avoir un impact sur l'apprentissage : modification des exemples pendant l'apprentissage (stacking), modification de l'importance relative des exemples (boosting) ; ou uniquement en post-traitement de l'apprentissage pour combiner les différents modèles appris (bagging).

4.1 Stacking

L'approche "Stacking" consiste à apprendre une succession de classifieurs organisés en cascade et tels que chaque classifieur va apprendre de nouveaux descripteurs permettant de redécrire les données.

L'une des illustrations de ce type d'approche est le réseau de neurones avec une couche cachée [Wol92]. La couche d'entrée contient autant de neurones que les données contiennent de descripteurs. La couche cachée est en général utilisée pour réduire le nombre de descripteurs par un mécanisme d'apprentissage. Ces nouveaux descripteurs sont ensuite fournis à la couche de sortie qui apprend à prédire les étiquettes associées aux exemples.

Aucune contrainte n'impose d'utiliser les mêmes algorithmes pour apprendre les différents classifieurs. Cette approche peut se généraliser avec n couches de stacking.

L'algorithme 1.1 présente un algorithme générique de Stacking comportant n_C couches, avec $f(c)$ classifieurs pour chaque couche c.

4.2 Boosting

Le Boosting, proposé par [FS97, Sch99], consiste à utiliser un algorithme d'apprentissage "faible" (dont le taux d'erreur est au moins inférieur à 50%) et d'effectuer T itérations d'apprentissage en se focalisant davantage, à chaque itération, sur les exemples précédemment mal appris.

Dans cette approche, un poids est associé à chaque exemple (initialement, distribution uniforme des poids), puis, à chaque nouvelle itération, un classifieur est appris à partir des données pondérées et l'erreur en prédiction associée à chaque exemple permet de réduire le poids des exemples correctement prédits tout en augmentant le poids des exemples incorrectement prédits. Une telle approche permet non seulement d'apprendre un modèle dont les performances sont souvent comparables aux meilleurs approches dédiées, mais elle permet également d'apprendre une pondération des exemples de l'ensemble d'entraînement qui met l'accent sur la difficulté d'être appris par les algorithmes d'apprentissage utilisés. Une telle information peut être utilisée par un expert du domaine pour analyser plus finement ces exemples et essayer de comprendre leurs spécificités.

L'algorithme 1.2 présente une forme générique du boosting.

4.3 Bagging

Enfin, le Bagging consiste à apprendre un ensemble de classifieurs à partir de différents tirages avec remise des données initiales. Chaque échantillon de données créé contient autant d'exemples que l'échantillon initial. Un échantillon est obtenu en tirant avec remise n exemples à partir de $\mathcal{A}$. Un tel échantillon

ALGORITHME 1.1: Algorithme générique de Stacking

Entrée :
les données d'apprentissage : $\mathcal{A} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$
n_C : nombre de couches
$f(c)$: nombre de classifieurs pour la couche c
$\mathcal{L}_1^c, \ldots, \mathcal{L}_{f(c)}^c$: classifieurs de la couche c
Sortie :
Méta-classifieur $\mathcal{H}$
début
  $\mathcal{A}^0 = \mathcal{A}$
  pour tous les *($c \in \{1 \ldots C\}$)* **faire**
    pour tous les *($j \in \{1 \ldots f(c)\}$)* **faire**
      $h_j^c = \mathcal{L}_j^c(\mathcal{A}^{c-1})$ *% Construction du $j^{ème}$ classifieur de la couche c*
    fin
    $\mathcal{A}^c = \emptyset$
    pour tous les *($i \in \{1 \ldots n\}$)* **faire**
      pour tous les *($j \in \{1 \ldots f(c)\}$)* **faire**
        $z_{ij} = h_j^c(x_i(\mathcal{A}^{c-1}))$ *% Calcul des nouveaux descripteurs à partir des données précédentes*
      fin
      $\mathcal{A}^c = \mathcal{A}^c \cup \{(z_{i1}, z_{i2}, \ldots, z_{if(c)}), y_i)\}$
    fin
  fin
  retourner $\mathcal{H}(\boldsymbol{x}) = h^c \circ h^{c-1} \circ \ldots h^1(\boldsymbol{x})$ *où h^c représente l'ensemble des classifieurs de la couche c et $\circ$ représente la composition des classifieurs après redescription des données par le classifieur précédent*
fin

ALGORITHME 1.2: Algorithme générique de Boosting

Entrée :
les données d'apprentissage : $\mathcal{A} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$
T : nombre d'itérations d'apprentissage
$\mathcal{L}$: classifieur de base
Sortie :
Méta-classifieur $\mathcal{H}$
début
% Initialisation uniforme des poids de chaque exemple
pour tous les *($i \in \{1 \ldots n\}$)* **faire**
$W_1(i) = \frac{1}{n}$
fin
pour tous les *($t \in \{1 \ldots T\}$)* **faire**
$h_t = \mathcal{L}(\mathcal{A}, W_t)$ *% Apprentissage du classifieur h_t à partir des données $\mathcal{A}$ et des poids W_t*
$\epsilon_t = \sum_{i=1}^{n} W_t(i) \times \mathbb{1}_{h_t(\boldsymbol{x}_i) \neq y_i}$ *% Calcul de l'erreur = somme des poids des exemples incorrectement prédits*
$\alpha_t = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon_t}$ *% Calcul du poids du classifieur h_t*
% Mise à jour des poids des exemples
% Z_t est un facteur de normalisation des poids
% Le poids des exemples correctement prédits à l'étape t va diminuer pour l'étape suivante
$W_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t h_t(\boldsymbol{x}_i) y_i}$
fin
retourner $\mathcal{H}(\boldsymbol{x}) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})\right)$
fin

est appelé un échantillon bootstrap. Lorsque tous les classifieurs ont été obtenus, ils sont combinés par une fonction d'agrégation. Dans le cas d'une prédiction binaire, cette fonction peut-être un simple vote majoritaire, un vote pondéré par l'erreur associée à chaque classifieur (erreur estimée sur les données non utilisées pour construire l'échantillon), ou tout autre fonction permettant d'agréger les prédictions de chaque classifieur.

ALGORITHME 1.3: Algorithme générique de Bagging avec aggrégation par vote majoritaire.

Entrée :
les données d'apprentissage : $\mathcal{A} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$
T : nombre d'itérations d'apprentissage
$\mathcal{L}$: classifieur de base
Sortie :
Méta-classifieur $\mathcal{H}$
début
  pour tous les $(t \in \{1 \ldots T\})$ **faire**
    $\mathcal{A}_t = Bootstrap(\mathcal{A})$ % *Création de l'échantillon par tirage avec remise à partir de* $\mathcal{A}$
    $h_t = \mathcal{L}(\mathcal{A}_t)$ % *Apprentissage du classifieur* h_t
  fin
  retourner $\mathcal{H}(\boldsymbol{x}) = \underset{y \in \mathcal{Y}}{argmax} \sum_{t=1}^{T} \mathbb{1}_{y=h_t(\boldsymbol{x})}$ % $\mathbb{1}_{expr}$ *renvoie* 1 *si* $expr$ *est évalué à vrai et* 0 *sinon*
fin

Dans le cadre de la prédiction des interactions physiques entre protéines (amarrage protéine-protéine), nous avons mis en place une stratégie proche du Bagging pour filtrer les conformations les moins plausibles. Cette stratégie exploite les prédictions négatives d'un ensemble de classifieurs et les seuls exemples rejetés sont tels qu'aucun classifieur ne les prédit positif. Les détails de cette approche sont présentés dans l'article [BBAP11].

Dans le cadre de la prédiction des interactions protéine-protéine directement à partir des séquences d'ADN, nous avons également mis en place une stratégie de type Bagging en adaptant le modèle de base de la manière suivante :

- compte tenu de la taille trop importante des données, les échantillons bootstrap sont de taille inférieure au nombre total d'exemples ;
- pour chaque échantillon bootstrap, un ensemble de classifieurs binaires est appris ;
- les prédictions de chaque classifieur sont combinées de manière à obtenir une fonction d'ordonnancement sur les données.

Nous avons également mis en place une stratégie visant à enrichir les données en ajoutant des descripteurs aux données initiales. Cette stratégie, proche de l'approche Stacking, consiste à associer à chaque exemple (ici, une paire de protéines) un vecteur de descripteurs dont chaque composante apporte une information sur la force du lien unissant les deux protéines. Ces descripteurs sont issus du domaine de l'apprentissage non-supervisé et plus particulièrement des mesures de qualité utilisées pour l'évaluation des règles d'association. Ces mesures sont présentées dans la section C.

Les détails de la mise en œuvre sont présentés dans le chapitre 2, ainsi que dans l'article [dVDA12].

C Apprentissage non-supervisé

Le paradigme de l'apprentissage non-supervisé consiste à rechercher des régularités dans des masses de données, en général non étiquetées.

L'extraction de régularités peut être divisée en deux catégories : le clustering et l'extraction de motifs fréquents.

1 Clustering

Le clustering permet de mettre en évidence une structuration interne des données. Cette structuration peut ne pas avoir été explicitée lors de la création des données. L'extraction de ce type d'information est réalisée à l'aide d'une métrique (distance ou similarité) permettant de comparer deux données (distance ou similarité), et d'un algorithme permettant de construire les clusters de données tels que la distance (resp. similarité) intra-cluster soit minimale (resp. maximale) et que la distance (resp. similarité) inter-cluster soit maximale (resp. minimale). Il existe de nombreux algorithmes : k-means, nuées dynamiques, classification hiérarchique ascendante, *etc*. Le lecteur intéressé pourra consulter [Ber02, JMF99] pour une présentation relativement exhaustive des différentes approches de clustering et [AZ12] pour un état de l'art récent sur le clustering en fouille de textes.

2 Extraction de motifs fréquents et règles d'association

L'extraction de motifs fréquents représente une autre stratégie de fouille des données non supervisée. Cette approche permet d'extraire, à partir de grandes masses de données essentiellement à valeurs binaires, des motifs (ensemble de descripteurs) fréquemment observés dans les données. L'un des algorithmes fondateurs de ce domaine est APRIORI proposé par [AIS93, AS94a].APRIORI est un algorithme permettant d'extraire à partir de **transactions** (i.e ensemble d'éléments) des régularités sous la forme de motifs (i.e des sous-ensembles d'éléments). Un motif est défini par l'ensemble des éléments qu'il contient et par l'ensemble des transactions dans lesquelles le motif apparaît. Pour de nombreuses applications, les motifs les plus fréquents sont recherchés.

Le **support** d'un motif correspond au nombre de transactions le contenant. Un motif est **fréquent** si son support est supérieur ou égal à un seuil de support minimal (ou de fréquence minimale) défini par l'utilisateur. Ainsi, si nous considérons l'ensemble des transactions suivant : $\{ABCD, ABDE, AD, BCD, BDE\}$ et le seuil de support minimal $3/5$ alors les motifs A, B, AD et BD sont fréquents. En effet, le motif AB apparaît 3 fois dans les transactions $ABCD$, $ABDE$, AD. Alors que le motif E n'apparaît que 2 fois dans les transactions $ABDE$ et BDE. Il est donc non fréquent. Depuis 1994, des représentations compactes de ces motifs, ainsi que les algorithmes associés pour les extraire, ont été proposés : les maximaux, les fermés, les libres, *etc*. Le lecteur intéressé pourra se référer à [Bor10] et chapitre 4 de [WKQ+08] pour des revues relativement récentes des méthodes d'extraction de motifs fréquents.

Bien que l'obtention de ces motifs fréquents représente un défi en soi, particulièrement lorsque les données sont denses et que le seuil de fréquence minimal est faible, notre objectif n'est pas ici de présenter ces méthodes mais plutôt une forme particulière de connaissance que nous pouvons en extraire : les **règles d'association**.

Une règle d'association peut être représentée sous la forme générale suivante : **prémisse** $\rightarrow$ **conclusion** où la prémisse et la conclusion représente des motifs fréquents dont l'intersection est vide.

[AIS93, AS94a] proposent d'évaluer l'intérêt des règles d'association avec deux mesures : **le support** défini précédemment et **la confiance**.

Le support correspond simplement au pourcentage d'exemples vérifiant la règle. La confiance correspond au pourcentage d'exemples pour lesquelles prémisse et conclusion sont vérifiées parmi les exemples pour lesquels la prémisse est vérifiée. Pour revenir à l'exemple précédent, les quatre règles d'association suivantes peuvent être extraites à partir des motifs fréquents :

- $A \rightarrow D$ (support = 3/5, confiance = 3/3)
- $D \rightarrow A$ (support = 3/5, confiance = 3/5)
- $B \rightarrow D$ (support = 4/5, confiance = 4/4)
- $D \rightarrow B$ (support = 4/5, confiance = 4/5)

Les règles $A \rightarrow D$ et $B \rightarrow D$ ont une confiance égale à 1 car il n'existe aucun exemple possédant la conclusion, mais pas la prémisse. Inversement, les règles $D \rightarrow A$ et $D \rightarrow B$ ont une confiance inférieure à 1 car il existe des exemples possédant la conclusion, mais ne possédant pas la prémisse (BCD et BDE pour la règle $D \rightarrow A$ et AD pour la règle $D \rightarrow B$).

La confiance permet donc d'évaluer la force du lien entre la prémisse et la conclusion de la règle d'association. Depuis 1994, de nombreux travaux ont montré que la confiance seule ne suffisait pas à évaluer finement l'intérêt des règles d'association [BMS97, LMP+03, LVL07, LT04a, LFZ99].

Dans la section suivante, nous allons présenter, de manière non exhaustive, différentes mesures utilisées pour évaluer l'intérêt des règles d'association.

3 Règles d'association et mesures de qualité

Le terme **mesure de qualité** est utilisé pour décrire les métriques permettant l'évaluation des règles d'association.

Comme nous l'avons vu, la confiance est l'une des mesures de qualité les plus simples à appréhender pour un non spécialiste de la fouille de données. En effet, elle peut s'interpréter comme la probabilité conditionnelle d'observer la conclusion sachant la prémisse. Elle est par nature même non symétrique car elle va potentiellement évaluer différemment les règles $A \rightarrow B$ et $B \rightarrow A$ où A et B sont deux motifs fréquents tels que $A \cap B = \emptyset$.

De nombreuses mesures ont été proposées pour évaluer la force du lien entre la prémisse et la conclusion d'une règle d'association. Parmi ces mesures, nous pouvons citer :

- **la nouveauté** [LFZ99] qui s'exprime comme $nouv(A \rightarrow B) = P(AB) - P(A)P(B)$. Si A et B sont indépendants alors le nombre d'exemples possédant à la fois A et B devrait être $P(A)P(B)$ or nous observons $P(AB)$. Cette mesure permet donc d'évaluer l'éloignement à l'indépendance entre A et B. Elle est symétrique.
- **le lift** [BMS97] qui s'exprime comme $lift(A \rightarrow B) = \frac{P(AB)}{P(A)P(B)}$. Le lift évolue entre 0 et $+\infty$. Si A et B sont indépendants alors le lift vaut 1. Si le lift vaut 2 alors A et B sont deux fois plus corrélés que sous l'hypothèse d'indépendance. Cette mesure est symétrique.
- **la moindre contradiction** (que j'ai proposé pendant ma thèse) [Azé03] qui se formule comme $MC(A \rightarrow B) = \frac{P(AB)-P(A\overline{B})}{P(B)}$ où $\overline{B}$ représente les exemples pour lesquels B n'est pas renseigné. Elle permet de mettre en évidence les règles d'association correspondant à des pépites de connaissances (peu de contre-exemples et conclusion avec faible support). Cette mesure est non-symétrique.

Dans le cadre de l'analyse des interactions protéine-protéine (présentées dans le chapitre 2, nous avons enrichi les données initiales par des descripteurs permettant d'évaluer la force du lien entre deux

protéines décrites par des caractéristiques binaires. Nous nous sommes focalisés sur l'utilisation de plusieurs mesures de qualité symétriques (évaluant de la même manière les règles $A \rightarrow B$ et $B \rightarrow A$) et sur des mesures non symétriques (telles que la confiance) que nous avons évaluées pour les deux formes de règles ($A \rightarrow B$ et $B \rightarrow A$).

Les détails de notre approche sont présentés dans le chapitre 2.

Sommaire

A Contexte général
1 Principe général de l'approche
2 Apprentissage du modèle
B Résultats
1 Prédiction des paires de protéines en interaction
2 Élagage des paire de protéines
C Conclusion

CHAPITRE

2

PRÉDICTION D'INTERACTIONS PROTÉINE-PROTÉINE

Principales publications du chapitre

[dVDA12] **Efficient Prediction of Co-Complexed Proteins Based on Coevolution**, PLoS One 2012

A Contexte général

Les interactions protéine-protéine sont impliquées dans la plupart des processus cellulaires. La connaissance du réseau d'interaction complet des interactions protéines d'un organisme donné (son **interactome**) facilite la compréhension des processus biologiques tels que les voies de signalisation, les voies métaboliques ou les mécanismes de transcription [PRJS05]. Cela permet aussi de prédire les fonctions de protéines non annotées à partir des fonctions des protéines avec lesquelles elles interagissent.

Les protéines en interaction, physique ou non, sont supposées co-évoluer. Ainsi, tout événement évolutif affectant une protéine peut avoir un impact direct ou indirect sur l'ensemble de ses interactants (voir [LR10] pour une revue récente). Cette hypothèse est à la base de plusieurs approches bioinformatiques ayant pour objectif de prédire l'ensemble des associations fonctionnelles entre protéines, i.e. les protéines qui font partie du même complexe cellulaire tel que défini par la biologie systémique.

Les deux principales approches de détection expérimentale des interactions protéine-protéine sont l'approche double-hybride et l'approche TAP-TAG.

- L'approche **double-hybride** est très utilisée pour la détection de complexes à grande échelle. Elle permet la détection indirecte de l'interaction qui est révélée par l'induction d'un gène rapporteur [FkS89]. Cette méthode est malheureusement relativement peu fiable car les taux de faux positifs et de faux négatifs sont relativement élevés. Cette méthode a été très largement utilisée pour détecter systématiquement les interactions protéine-protéine chez la levure.
- L'approche **TAP-TAG** dont le principe consiste à produire une protéine appât en fusion avec deux étiquettes de purification, généralement la calmoduline et la protéine A. La purification sur colonne d'affinité permet de retenir la protéine appât, qui est liée à ses partenaires cellulaires. Le complexe est ensuite élué (décroché de la colonne d'affinité), puis les composants sont séparés et identifiés par spectroscopie de masse.

Malheureusement, comme le montre [ITM+00], chaque méthode tend à découvrir des interactions non détectées par les autres méthodes. Il est donc difficile d'évaluer la "fiabilité" des interactions détectées si seule une méthode les détecte.

Parmi les méthodes de prédiction *in silico*, nous nous focaliserons sur les méthodes exploitant les profils phylogénétiques et le contexte génomique.

- L'approche **Phylogenetic Profiles** (PP, [PMT+99]) explore les motifs de type "présence/absence" de protéines dans un ensemble d'espèces donné. L'hypothèse est que deux protéines sont en interaction pour réaliser une fonction particulière. Si une espèce perd, à la suite d'un événement évolutif (mutation, croisement, *etc*), une des deux protéines alors la fonction disparaît et l'autre protéine impliquée dans cette fonction devrait également disparaître, ce qui conduira, après quelques événements évolutifs, à des profils phylogénétiques comparables.
- L'approche **Genomic Context** (GC, [DSHB98]) explore la conservation du voisinage des gènes dans différentes espèces. Si deux protéines sont en interaction, alors la proximité physique des deux gènes associés est supposée être conservée dans les différentes espèces concernées.
- Et enfin l'approche **mirrortree** ([PV01]) dans laquelle des arbres phylogénétiques sont construits à partir d'alignements multiples. Des paires d'arbres sont ensuite comparées. Les arbres ayant une forte similarité sont supposés représenter des paires de protéines en interaction alors que des arbres présentant une faible similarité sont associés à des protéines n'interagissant pas.

Depuis quelques années, plusieurs approches exploitant des techniques d'apprentissage sont apparues dans la littérature, principalement pour l'identification d'interactions protéine-protéine dans la levure. Ces méthodes exploitent massivement l'intégration de données multiples issues de sources très

hétérogènes, y compris des données expérimentales : séquences des protéines, prédiction d'interactions protéine-protéine obtenues à partir d'expériences à haut débit, données d'expression génique, termes GO, données de co-régulation, données de localisation, fluctuation de l'expression des micro-ARN dans le cycle de vie de la levure, *etc* [JYG+03, LWJ+04, LXP+05, QBJKS06, QN08]. Ces méthodes fournissent d'excellents résultats, même si un nombre restreint de descripteurs est pris en considération pour la prédiction [LXP+05]. Cependant, compte tenu du fait que ces méthodes nécessitent des sources d'informations variées et notamment des données expérimentales, leur utilisation est restreinte à *S. cerevisiae* car de telles données sont très rarement disponibles pour d'autres organismes.

Récemment, [GJJE+10] a proposé une nouvelle méthode d'apprentissage pour la détection d'interactions protéine-protéine reposant sur l'utilisation de données issues de plusieurs méthodes de prédictions différentes qui ont été conçues indépendamment (PP, GC, I2H, *mirrortree* et la fusion de gènes (GF)). Même si cette nouvelle approche appliquée à *E. coli* fournit de meilleurs résultats que chacune des méthodes prises indépendamment, le nombre de faux positifs (FP) et de faux négatifs (FN) demeure élevé. De plus, comme les données utilisées par les différentes méthodes combinées par [GJJE+10] sont de natures différentes et fournies par des approches indépendantes, le temps de calcul global de l'intégralité des données est élevé et la quantité de données pour lequelles une partie des informations ne peut pas être renseignée est souvent non nulle. Ceci conduit donc inévitablement à l'introduction de valeurs manquantes dans la description des données. La prise en considération de ces valeurs manquantes est souvent non triviale.

Avec Damien de Vienne (post-doc au Centre for Genomic Regulation (CRG, Barcelona)), nous avons proposé une nouvelle approche pour la détection des protéines appartenant à des complexes impliqués dans des fonctions cellulaires spécifiques [dVDA12]. Cette méthode repose uniquement sur les séquences des protéines et ne nécessite donc aucune donnée expérimentale. L'ensemble des caractéristiques liées à la coévolution des protéines est extrait puis, une approche de type Ensemble Learning est mise en œuvre pour combiner toutes ces informations dans un modèle prédictif.

Nous avons étendu l'approche PP avec des descripteurs classiquement utilisés en apprentissage non supervisé pour évaluer les règles d'association (voir chapitre 1, section C). Étant donné une paire de protéines (p_A, p_B), le calcul de ces nouveaux descripteurs repose uniquement sur le nombre d'espèces ayant un orthologue[7] pour p_A, le nombre d'espèces ayant un orthologue pour p_B, le nombre d'espèces ayant un orthologue pour p_A et p_B, ainsi que le nombre total d'espèces considéré.

La figure 2.1 présente le protocole mis en place pour calculer les descripteurs utilisés afin d'apprendre un modèle permettant de détecter les interactions protéine-protéine.

Nous avons appliqué notre méthode à la prédiction des interactions protéine-protéine dans *E. coli*, organisme très largement étudié et dont l'interactome est bien défini. Nous avons extrait les paires de protéines en interaction à partir de la base de données Ecid [LEG+09]. La comparaison des résultats obtenus avec les méthodes classiques nous permet de conclure que l'approche que nous avons proposée est très efficace compte tenu de la nature des descripteurs utilisés.

Nous présentons, dans la suite de ce document, les données utilisées, le cœur de l'apprentissage, ainsi que quelques résultats. Pour plus de détails, le lecteur intéressé pourra se référer à [dVDA12].

[7] Deux gènes (protéines) présents dans deux espèces différentes sont dits orthologues s'ils ont évolué depuis une unique séquence présente dans le dernier ancêtre commun aux deux espèces.

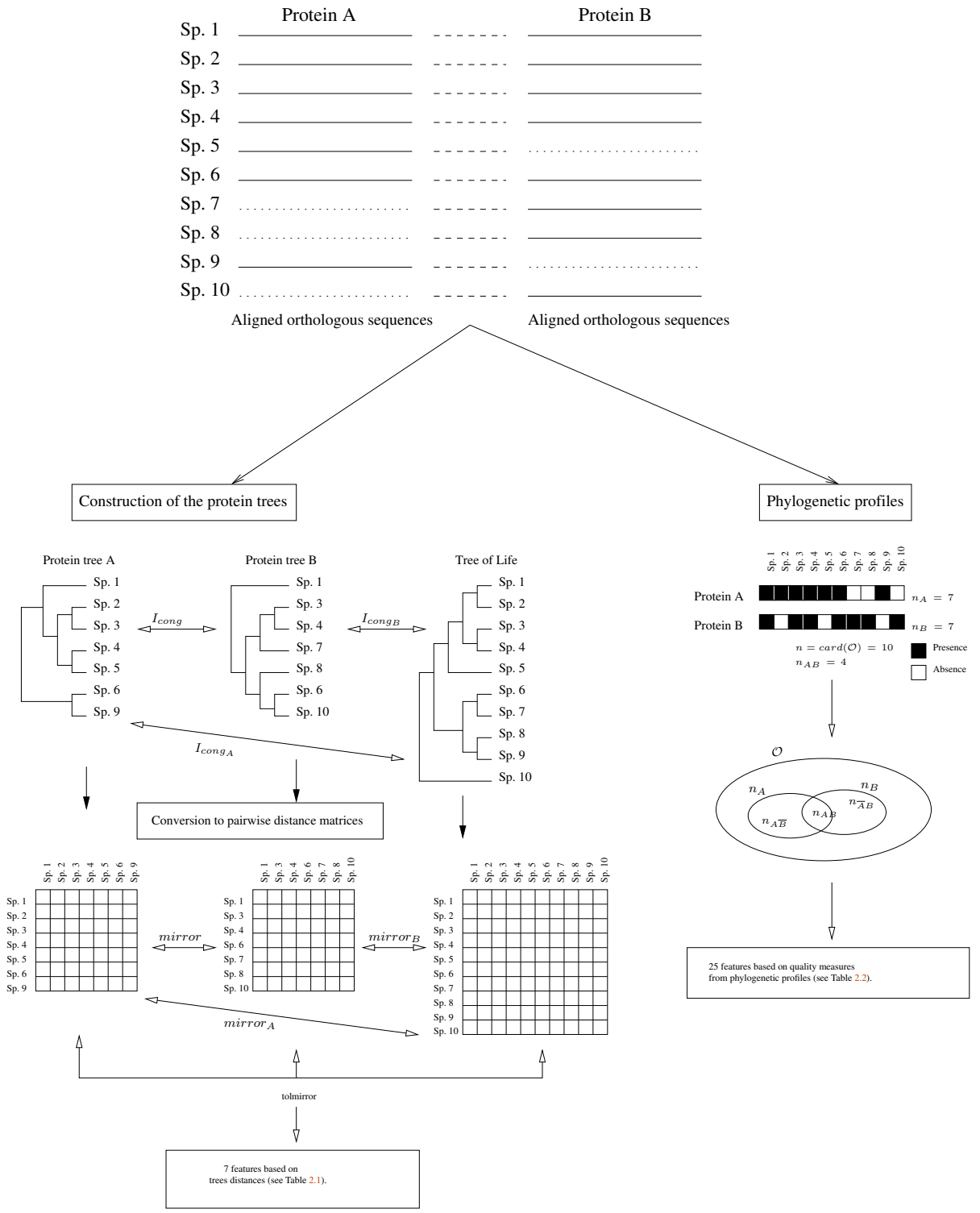

FIGURE 2.1 : Principe du calcul des descripteurs utiles pour apprendre un modèle permettant de prédire les interactions protéine-protéine.

TABLE 2.1 : Liste des descripteurs basés sur le contexte génomique

Name	Description/formula	References
$mirror$	Tree similarity between proteins A and B computed as the correlation between their pairwise distance matrices	[GBJ+00, PV01]
$mirror_A$	Tree similarity between protein A and the Tree of Life (ToL) with the mirror method	[GBJ+00, PV01]
$mirror_B$	Tree similarity between protein B and the ToL with the mirror method	[GBJ+00, PV01]
$tol - mirror$	Tree similarity between proteins A and B based on the mirror method after correction of their pairwise distance matrices to remove the background similarity due to speciation of the species themselves	[PRJS05]
I_{cong}	Topological similarity between the trees of proteins A and B as estimated by the size of the maximum agreement subtree (MAST) between the two trees	[dVGM07]
$I_{cong}A$	Topological similarity (I_{cong} index) between the tree of protein A and the ToL	[dVGM07]
$I_{cong}B$	Topological similarity (I_{cong} index) between the tree of protein B and the ToL	[dVGM07]

1 Principe général de l'approche

Pour chaque protéine du génome d'*E. coli*, les séquences orthologues, dans 115 autres génomes prokaryotes, sont calculés ([JPV08]), puis alignées (2 177 alignements multiples). Le profil phylogénétique d'une protéine est caractérisé par le vecteur des espèces pour lesquelles la protéine possède un orthologue. Les paires de protéines sont ensuite évaluées en comparant soit leurs profils phylogénétiques (partie droite de la figure 2.1), soit en comparant leurs arbres phylogénétiques (partie gauche de la figure 2.1). À partir de ces deux comparaisons, un ensemble de 35 descripteurs numériques est calculé (voir 2.1 et 2.2). Ces descripteurs numériques sont alors utilisés pour apprendre un modèle permettant d'ordonner au mieux l'ensemble des paires de protéines. Ce modèle est obtenu par combinaison de classifieurs binaires.

Afin de pouvoir comparer les performances obtenues avec les approches existantes, nous avons utilisé des métriques standards obtenues à partir de la liste des paires de protéines triées par valeurs décroissantes du score calculé par le modèle appris. Ces métriques sont la précision, le rappel, l'aire sous la courbe ROC, ainsi que des indicateurs graphiques tels que la courbe ROC.

2 Apprentissage du modèle

Nous avons utilisé des approches très "classiques" en apprentissage supervisé dans leur version implantées dans Weka [HFH+09]. Deux types d'arbre de décision ont été utilisés (J48 et RandomForest), ainsi que deux types de règles de décision (PART et JRip). Nous avons également utilisé la version "bagged" de ces quatre classifieurs. Les prédictions obtenues par ces 8 classifieurs ont ensuite été combinées pour obtenir un score permettant d'ordonner les paires de protéines.

Nous nous situons ici dans le cadre général de l'ensemble learning tel que présenté dans le chapitre

TABLE 2.2 : Liste des descripteurs basés sur les profils phylogénétiques.

Name	Description/formula	References
n_A, n_B and n_{AB}		
confidence	$\frac{n_{AB}}{n_A}$	[AS94b]
recall	$\frac{n_{AB}}{n_B}$	[LFZ99]
lift	$\frac{nn_{AB}}{n_A n_B}$	[BMS97]
dice	$\frac{2 \times n_{AB}}{n_A + n_B}$	[Dic45]
pearson	$\frac{nn_{AB} - n_A n_B}{\sqrt{n_A n_B n_{\overline{A}} n_{\overline{B}}}}$	[Pea00]
GI	$log\left(\frac{n_{AB} n}{n_A n_B}\right)$	[CH90]
IQC	$2 \times \frac{P(AB) - P(A)P(B)}{P(A)P(\overline{B}) + P(\overline{A})P(B)}$	[Coh60]
confidenceCentered 1	$\frac{nn_{AB} - n_A n_B}{nn_A}$	[LT04b]
confidenceCentered 2	$\frac{nn_{AB} - n_A n_B}{nn_B}$	[LT04b]
leastContradiction 1	$\frac{n_{AB} - n_{A\overline{B}}}{n_B}$	[AK02]
leastContradiction 2	$\frac{n_{AB} - n_{\overline{A}B}}{n_A}$	[AK02]
jaccard 1	$\frac{n_{AB}}{n_{A\overline{B}} + n_B}$	[Jac08]
jaccard 2	$\frac{n_{AB}}{n_{\overline{A}B} + n_A}$	[Jac08]
loevinger 1	$1 - \frac{nn_{A\overline{B}}}{n_A n_{\overline{B}}}$	[Loe47]
loevinger 2	$1 - \frac{nn_{\overline{A}B}}{n_{\overline{A}} n_B}$	[Loe47]
tec 1	$\frac{n_{AB} - n_{A\overline{B}}}{n_{AB}}$	
tec 2	$\frac{n_{AB} - n_{\overline{A}B}}{n_{AB}}$	
LAP 1	$\frac{n_{AB} + 1}{n_A + 2}$	[Goo03]
LAP 2	$\frac{n_{AB} + 1}{n_B + 2}$	[Goo03]
GAN 1	$\frac{2 * n_{AB}}{n_A} - 1$	[Gan87]
GAN 2	$\frac{2 * n_{AB}}{n_B} - 1$	[Gan87]
Zhang 1	$\frac{P(AB) - P(A) \times P(B)}{max\left(P(AB) \times P(\overline{B}), P(A\overline{B}) \times P(B)\right)}$	[Zha00]
Zhang 2	$\frac{P(AB) - P(A) \times P(B)}{max\left(P(AB) \times P(\overline{A}), P(\overline{A}B) \times P(A)\right)}$	[Zha00]
Pearl 1	$P(AB) \times \left\lvert \frac{P(AB)}{P(A) - P(B)} \right\rvert$	[Pea88]
Pearl 2	$P(AB) \times \left\lvert \frac{P(AB)}{P(B) - P(A)} \right\rvert$	[Pea88]

1 et plus particulièrement dans le cadre du Bagging où la fonction de vote est construite à partir des prédictions binaires de chacun des classifieurs mais également à partir de la "confiance" associée à chaque prédiction. Ainsi, le score associé à chaque paire x du jeu de test est obtenu de la manière suivante :

$$S(x) = S_{pos}(x)/S_{neg}(x)$$

où $S_{pos}(x)$ (resp. $S_{neg}(x)$) représente le score associé à x pour la classe **pos** (resp. **neg**).

Les scores $S_{pos}(x)$ et $_{neg}(x)$ sont calculés de la manière suivante :

$$S_{pos}(x) = \begin{cases} P_{pos} \times e^{n_{pos}} & \text{si } n_{pos} > 0 \\ 1 & \text{sinon} \end{cases}$$

$$S_{neg}(x) = \begin{cases} P_{neg} \times e^{n_{neg}} & \text{si } n_{neg} > 0 \\ 1 & \text{sinon} \end{cases}$$

$$\text{avec} \begin{cases} n_{pos} = \sum_{c \in C} \mathbb{1}_{P_c(x) \geq 0.5} \\ P_{pos} = \prod_{c \in C} P_c(x) \times \mathbb{1}_{P_c(x) \geq 0.5} \end{cases}$$

$$\text{et} \begin{cases} n_{neg} = \sum_{c \in C} \mathbb{1}_{P_c(x) < 0.5} \\ P_{neg} = \prod_{c \in C} (1 - P_c(x)) \times \mathbb{1}_{P_c(x) < 0.5} \end{cases}$$

où C représente l'ensemble des classifieurs et $P_c(x)$ représente la "confiance" associée par chaque classifieur à la prédiction pos ; n_{pos} (resp. n_{neg}) représente le nombre de classifieurs ayant prédit la classe pos (resp. neg).

L'ensemble des résultats a été obtenu selon un protocole de 3 validations croisées. Pour éviter tout biais lié à l'utilisation de la validation-croisée, le protocole a été itéré 30 fois. Les résultats présentés dans la section suivante correspondent aux valeurs moyennes observées pour chaque métrique, ainsi qu'aux valeurs minimales et maximales associées à ces mêmes métriques.

B Résultats

Les résultats sont présentés avec les critères d'évaluation suivants : courbe ROC, aire sous la courbe ROC et courbes précision-rappel en fonction du classement des paires de protéines. Les résultats obtenus avec les méthodes classiques sont présentés : *mirrortree* [PV01], *tol-mirror* [CL07, PRJS05, SYKT05] et *context-mirror* [JPV08] ainsi que ceux obtenus avec notre méthode [dVDA12]. Seul un sous-ensemble des résultats obtenus est présenté dans ce document. Pour plus de détails, le lecteur intéressé pourra se reporter à la publication associée [dVDA12].

1 Prédiction des paires de protéines en interaction

La figure 2.2 présente les performances comparées des 4 approches que nous avons testées. Nous pouvons voir que l'aire sous la courbe ROC associée à notre approche (voir figure 2.2,A pour la comparaison des courbes ROC et 2.2,B pour la comparaison des aires) est égale en moyenne à 0.93 avec une variance inférieure à 3.10^{-3} sur l'ensemble des 30 modèles appris indépendamment. L'aire sous la courbe ROC associée à la meilleure approche concurrente (*context mirror*) est égale à 0.868 (voir figure 2.2,C).

Si nous nous focalisons sur les 200 premières paires de protéines prédites comme étant en interaction, la précision obtenue par l'approche *context mirror* sur ces 200 paires est de 28.5% (avec une précision égale à 100% pour les 13 premières paires), alors que notre approche permet d'obtenir une précision moyenne de 95.5% sur les 200 premières paires ordonnées (avec une précision égale à 100% pour les 90 premières paires).

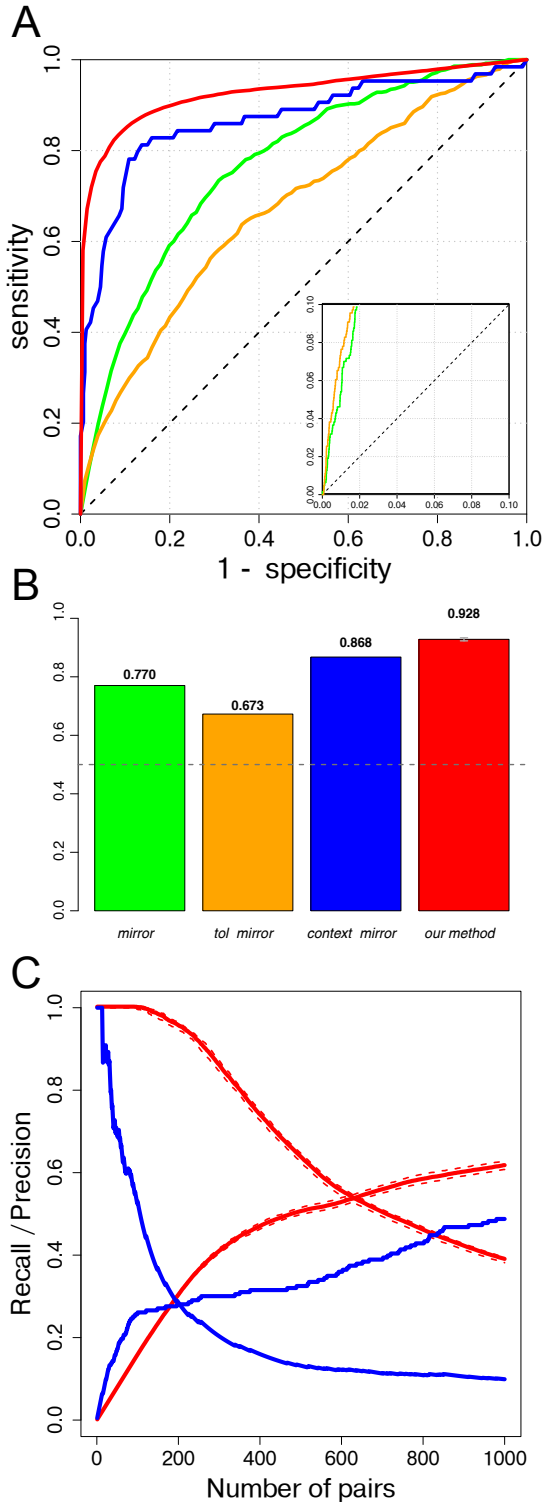

FIGURE 2.2 : Comparaison des performances des approches *mirrortree* [PV01], *tol-mirror* [CL07, PRJS05, SYKT05] et *context-mirror* [JPV08] avec notre méthode [dVDA12]

La différence notable de performances entre notre approche et l'approche *context mirror* est principalement due à deux éléments : (i) nous utilisons plusieurs descripteurs dont la combinaison s'avère très efficace et (ii) nous apprenons de manière supervisée un modèle prédictif à partir de données de bonne qualité.

De plus, l'utilisation d'une approche de type ensemble learning nous permet non seulement d'obtenir une fonction d'ordonnancement pour les paires de protéines, mais également une fonction de filtre basée sur un consensus de prédictions négatives.

2 Élagage des paire de protéines

Nous avons conçu et testé une procédure d'élagage des paires de protéines reposant sur le principe suivant : (i) consensus total des classifieurs pour prédire la classe négative, (ii) confiances élevées des différents classifieurs dans la prédiction négative (ce qui se traduit par une confiance dans la prédiction négative, P_{neg}, proche de 1).

Nous avons également étudié l'impact de différents sous-ensembles des descripteurs disponibles pour apprendre le modèle prédictif. Les descripteurs utilisés peuvent être regroupés en cinq familles :

- **tree** restriction aux descripteurs obtenus à partir des arbres phylogénétiques. Cette famille de descripeurs est divisée en deux sous-familles :
  - **matrix** restriction aux descripteurs $mirror$ et $tol - mirror$
  - **topology** restriction aux descripteurs I_{cong}
- **PP** restriction aux descripteurs obtenus à partir des profils phylogénétiques
- **ALL** utilisation de l'ensemble des descripteurs

La figure 2.3 présente l'impact d'un seuil d'élagage associé à P_{neg} sur l'aire sous la courbe ROC.

Nous pouvons constater que l'utilisation de l'ensemble des descripteurs permet d'obtenir les meilleurs résultats.

Cependant, si nous nous plaçons dans un contexte "haut débit" où le temps de calcul doit être optimisé face à la masse de données à traiter, l'utilisation des descripteurs liés aux arbres phylogénétiques peut s'avérer problématique car coûteux à calculer. L'utilisation du sous-ensemble de descripteurs réduits aux seuls descripteurs obtenus à partir des profils phylogénétiques (**PP**) permet d'obtenir des performances largement comparables aux meilleures approches actuelles.

C Conclusion

L'approche que nous avons proposée avec Damien de Vienne permet d'obtenir de meilleurs résultats que les approches actuelles pour la prédiction d'interactions protéine-protéine uniquement basée sur les séquences des protéines.

Notre approche permet non seulement d'ordonner efficacement les paires de protéines par "probabilité" décroissante d'être réellement en interaction, mais elle permet également, via l'utilisation d'un filtre, de supprimer pratiquement toutes les paires négatives, en conservant environ la moitié des paires positives.

Ce résultat nous permet d'envisager l'utilisation d'une telle approche de détection des interactions protéine-protéine dans un cadre de prédiction systématique de l'interaction physique entre protéines ("cross-docking") où il est important de limiter au maximum le nombre d'amarrages protéine-protéine à effectuer.

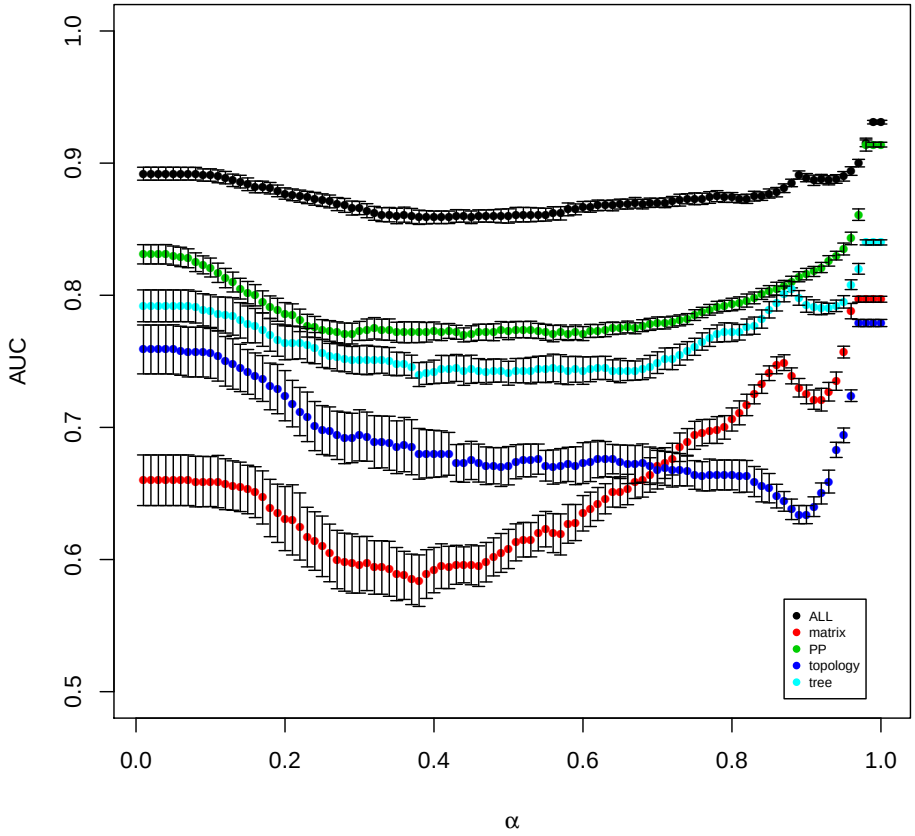

FIGURE 2.3 : Impact du seuil d'élagage sur l'AUC pour les 5 familles de descripteurs. Le seuil d'élagage α n'est appliqué que sur les paires de protéines pour lesquelles l'intégralité des classifieurs ont voté neg.

Nous présentons dans le chapitre suivant une approche permettant d'effectuer de l'amarrage protéine-protéine avec une précision et des temps de calcul tels que la combinaison avec l'approche de prédiction d'interactions protéine-protéine présentée dans ce chapitre rend envisageable l'analyse détaillée des interactions protéine-protéine d'un génome complet en mode "haut-débit".

Parmi les perspectives à court terme associées à ces travaux, dans le cadre d'une collaboration avec l'IGM d'Orsay, nous cherchons à prédire les interactions protéine-protéine de champignons. Nous disposons actuellement d'un ensemble de 48 champignons avec leurs orthologues, et cet ensemble devrait rapidement être étendu à plus de 150 champignons. Bien que n'étant pas exactement dans le même cadre que pour *E. coli* qui appartient à la famille des procaryotes alors que les champignons sont des eucaryotes, nous allons tester notre approche sur ces champignons.

Une autre perspective de travail à court terme consisterait, pour l'ensemble des complexes de *E. coli* dont la structure 3D est disponible, à appliquer notre approche sur l'élagage de la matrice de cross-docking, puis à réaliser les amarrages les plus probables afin de vérifier si les prédictions sont fiables ou non.

Notre approche pourrait également être appliquée pour prédire les interactions entre protéines d'un

génome mutant étant donné les interactions connues dans le génome de référence.

Enfin, compte tenu du coût décroissant du séquençage, de plus en plus de projets sont liés au séquençage d'un ensemble d'organismes proches d'un organisme de référence. Nous pourrions tester notre approche sur ce type de données.

Sommaire

A Introduction
B Modélisation atomique *vs* modélisation gros-grain
1 Modélisation gros-grain par diagramme de Voronoï et apprentissage de fonctions de score
2 Génération des conformations à l'aide des diagrammes de Voronoï
3 Génération : Hex, évaluation : Voronoï + apprentissage
C Compétition CAPRI
1 Critères d'évaluation de CAPRI
2 Résultats
D Conclusion

CHAPITRE

3

AMARRAGE PROTÉINE-PROTÉINE

Principales publications du chapitre

[BBAP11] **A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions**, PLoS One 2011
[BAPR11] **Amarrage protéine-protéine par couplage de la complémentarité de forme et des empreintes Voronoï**, Jobim 2011
[ABH+11] **Using Kendall-Tau Meta-Bagging to Improve Protein-Protein Docking Predictions**, PRIB 2011
[BBAP09] **Comparing Voronoi and Laguerre tessellations in the protein-protein docking context**, ISVD 2009
[BAJP07] **A new protein-protein docking scoring function based on interface residue properties**, Bioinformatics 2007
[BPAJ05] **A docking analysis of the statistical physics of protein-protein recognition**, Physical Biology 2005

A Introduction

L'intégration des signaux extra-cellulaires en une réponse biologique adaptée repose en grande partie sur l'association de complexes protéine-protéine. La détection et la détermination de l'organisation structurale de ces assemblages moléculaires représente donc une étape essentielle pour la compréhension de ces mécanismes et de leur régulation. Si les techniques qui permettent la détermination expérimentale des structures protéiques ont connu des avancées fondamentales, notamment grâce aux projets de génomique structurale, cette détermination reste délicate voire impossible, surtout lorsque l'objet étudié est un complexe (assemblage de plusieurs protéines). De plus, il a été démontré expérimentalement que le nombre de complexes existant *in vivo* était bien supérieur au nombre de protéines, rendant inenvisageable le recours systématique à l'expérimentation. L'amarrage protéine-protéine, qui consiste à prédire la structure tridimensionnelle de ces assemblages macromoléculaires à partir des structures des partenaires isolés, représente donc un outil crucial dans l'étude du fonctionnement de la cellule et est devenu un défi en biologie computationnelle [MGC+05, Rit08, YAR11].

L'un des objectifs qui anime la communauté travaillant sur cette problématique est de concevoir une méthode de prédiction des structures 3D permettant d'être réellement utilisé par les biologistes, par exemple dans le cadre d'application de "drug design". Pour arriver à atteindre cet objectif, la précision associée aux prédictions obtenues par une telle méthode doit donc être très élevée afin de minimiser le nombre de faux positifs.

En effet, les expériences biologiques sont très exigeantes et ne peuvent pas être multipliées car très coûteuses en temps et en argent. Par conséquent, une méthode de prédiction de structure 3D de complexes protéine-protéine ne doit fournir qu'un nombre très limité de structures possibles pour un complexe donné. De plus, une telle méthode doit pouvoir évaluer la probabilité qu'au moins une structure correcte soit présente dans l'ensemble des structures proposées afin de garantir aux biologistes qu'au moins une expérience biologique sera couronnée de succès.

Les différentes procédures existantes traitent généralement le problème en deux étapes :
- une première phase au cours de laquelle un grand nombre de conformations est généré (étape généralement la plus limitante en temps de calcul),
- puis une seconde phase au cours de laquelle ces différentes conformations sont évaluées afin d'extraire le sous-ensemble de conformations les plus proches possibles de la conformation native. Dans la suite du document, cet ensemble de conformations sera désigné par *l'ensemble des conformations quasi-natives*.

Certaines méthodes combinent ces deux phases pour limiter le nombre de conformations générées en garantissant que chaque conformation créée optimise au moins l'un des critères d'évaluation, compte tenu de l'ensemble des conformations déjà créées.

Parmi les différentes approches d'amarrage protéine-protéine existantes, nous distinguerons deux types de modélisations différentes des complexes : modélisation au niveau atomique *vs* modélisation gros grain.

B Modélisation atomique *vs* modélisation gros-grain

Compte tenu du nombre élevé d'atomes d'une protéine de taille moyenne, et du nombre de degrés de liberté devant être pris en considération pour chaque atome, la modélisation atomique des protéines pour la prédiction de l'amarrage protéine-protéine est très coûteuse. Cependant, cette modélisation permet d'obtenir une granularité telle qu'il devient possible de prendre en considération de très

nombreuses déformations des protéines lors de l'interaction. Bien que le temps de calcul associé à cette représentation soit important, plusieurs approches travaillent au niveau atomique avec notamment des fonctions énergétiques bien définies qui permettent de réduire le nombre de conformations à explorer [Gra06, GMW+03, DMS+05, WSFB05].

À l'opposé, la modélisation gros-grain des protéines, souvent effectuée au niveau des acides aminés, permet de réduire significativement la quantité de conformations à explorer. Mais ce gain en temps de calcul, s'accompagne généralement d'une dégradation dans la capacité à capturer des déformations importantes des protéines lors de l'interaction [LV11, WS11, Toz05, RE10, GMCF10]

Certaines approches telles que l'approche RosettaDock [LG08] combinent une étape de modélisation gros-grain suivie d'une étape de modélisation atomique.

L'introduction du récent article de Moal *et al.* [MAB11] dresse un panorama relativement exhaustif des différentes approches de prédiction d'affinité pour l'amarrage protéine-protéine.

Dans le cadre de nos travaux, nous avons contribué à la modélisation gros-grain des protéines par diagramme de Voronoï, aussi bien dans le cadre de l'évaluation des conformations, que de la génération des conformations. Nous présentons dans les sections suivantes nos contributions pour ces deux étapes clés de l'amarrage protéine-protéine.

1 Modélisation gros-grain par diagramme de Voronoï et apprentissage de fonctions de score

Dans le cadre de la thèse de Julie Bernauer [Ber06], nous avons testé l'hypothèse qu'une modélisation gros grain à l'aide de diagrammes de Voronoï [Pou04] couplée à un apprentissage supervisé de modèles prédictifs permettrait de différencier efficacement les quasi-natifs des leurres.

Cette approche se situe dans le cadre de la seconde phase "classique" des algorithmes d'amarrage protéine-protéine : évaluation d'un large ensemble de conformations. Cet ensemble de conformations a été obtenu de la façon suivante :

- collecte des natifs à partir de la Protein Data Bank (PDB[8], [BWF+00] ;
- création de quasi-natifs à partir des natifs en ajoutant de faibles perturbations autour de la position native ;
- génération des leurres en utilisant un algorithme d'amarrage [CDJ91] et en sélectionnant *a posteriori* les conformations suffisamment éloignées des natifs.

Les conformations sont ensuite modélisées par des diagrammes de Voronoï. Puis, pour chaque conformation, un ensemble de descripteurs géométriques est calculé. Enfin, une étape d'apprentissage supervisé est effectuée, à partir des conformations décrites par les descripteurs issus des diagrammes de Voronoï.

L'apprentissage supervisé était réalisé par un algorithme évolutionnaire dont la fonction objectif était de maximiser l'aire sous la courbe ROC. L'algorithme évolutionnaire estimait les poids w_i et valeurs de centrage c_i de fonction non linéaire suivante $f(X) = \sum_i w_i|X_i - c_i|$ où X représente un complexe, X_i la valeur du $i^{\text{ème}}$ descripteur de ce complexe.

Nous avons montré que la représentation gros grain des complexes par des diagrammes de Voronoï permettait d'apprendre des modèles efficaces pour séparer les natifs (et quasi-natifs) des leurres. La comparaison des performances avec des approches classiques en apprentissage supervisé (SVM, régression logistique, *etc*) a montré que l'approche par algorithme évolutionnaire offre de bien meilleures performances, à la fois en temps de calcul et en sensibilité et spécificité.

[8] `http ://www.pdb.org`

La comparaison de cette approche avec d'autres approches de Docking, notamment dans le cadre de la compétition CAPRI (voir section C), a également montré que cette approche permet d'obtenir des résultats concurrentiels.

Pour plus de détails, le lecteur intéressé pourra se reporter aux publications suivantes [BAJP05, BPAJ05, Ber06, BAJP07].

2 Génération des conformations à l'aide des diagrammes de Voronoï

Ayant montré dans nos précédents travaux que la modélisation des complexes protéine-protéine par des diagrammes de Voronoï permettait de différencier efficacement les natifs et quasi-natifs des leurres, nous nous sommes ensuite focalisés sur la première phase de l'amarrage protéine-protéine : la génération des conformations directement en gros-grain.

Ainsi, dans le cadre de la thèse de Thomas Bourquard [Bou09], nous avons participé à la conception d'une approche permettant d'exploiter la modélisation de Voronoï pour générer des conformations. Ces conformations étant ensuite évaluées via un modèle appris par un algorithme évolutionnaire comparable à celui décrit précédemment. Les résultats obtenus ont confirmé les précédents résultats obtenus lors de la thèse de Julie Bernauer [BBAP09, BBAP11].

Cependant, les temps de calculs associés à cette phase de génération des conformations sont trop importants pour pouvoir envisager une exploitation de ces approches en mode "haut-débit".

Dans le cadre d'un séjour post-doctoral effectué par Thomas Bourquard chez Dave Ritchie, nous avons pu entamer une collaboration fructueuse en combinant Hex [RK00] avec notre phase d'apprentissage de modèles prédictifs basés sur la représentation de Voronoï des complexes.

L'utilisation de l'algorithme de complémentarité de formes Hex implanté sur cartes graphiques (GPU) a permis de réduire considérablement le temps nécessaire pour l'échantillonnage statistique des quelques 10^9 modes d'associations possibles pour deux protéines de taille moyenne [MKTE01]. Cet algorithme est capable de générer et évaluer en quelques secondes plusieurs millions de conformations candidates afin d'en extraire un ensemble réduit de conformations d'intérêt [RV10]. Cependant la fonction d'évaluation intégrée dans Hex ne permet pas d'identifier de manière fiable une solution quasi-native dans cet ensemble, d'où le post-traitement que nous effectuons sur les conformations générées par Hex.

3 Génération : Hex, évaluation : Voronoï + apprentissage

Le post-traitement effectué sur les conformations obtenues par Hex est similaire à celui que nous effectuions dans le cadre des travaux de thèse de Julie Bernaueur. L'objectif est de pouvoir calculer l'ensemble des descripteurs géométriques obtenus à partir de la modélisation par diagrammes de Voronoï de chaque complexe. Pour obtenir ces informations, nous procédons en 3 étapes :

1. Plonger le complexe (ainsi que les deux protéines isolées) dans une cage de solvant artificiel ;
2. Calculer le diagramme de Voronoï de chaque protéine (uniquement pour les résidus accessibles au solvant). Ce calcul est réalisé en utilisant la librairie [BDP+02] qui permet de faire ces calculs en temps minimal (complexité $O(nlog(n))$ où n représente le nombre de résidus considérés comme accessibles au solvant) ;
3. Puis calculer les descripteurs géométriques permettant de caractériser le complexe.

Le modèle prédictif que nous utilisons est indépendant de Hex et repose sur le même principe que celui mis en place dans la thèse de Julie Bernauer et amélioré dans la thèse de Thomas Bourquard. Nous

avons testé différentes approches pour apprendre un modèle prédictif :
- algorithmes évolutionnaires, SVM, régression logistique [BAJP07] ;
- combinaisons de classifieurs [BBAP11] ;
- recherche d'un tri consensus [ABH+11].

Actuellement, nous obtenons les meilleurs résultats avec des approches de combinaisons de classifieurs ("ensemble learning") dont les propriétés ont été décrites dans le chapitre 1.

C Compétition CAPRI

De nombreuses équipes participent à la compétition internationale CAPRI[9] qui permet régulièrement aux équipes travaillant sur la problématique de l'amarrage protéine-protéine de s'évaluer. Le principe de cette compétition est le suivant :
- une ou plusieurs nouvelles structures de complexes sont résolues par une ou plusieurs équipes de chercheurs
- avant de rendre ces structures publiques, ces équipes les envoient aux organisateurs de CAPRI qui organise alors un nouveau "round" avec autant de "target" que de nouvelles structures. Chaque round/target se déroule en deux phases :
  1. une première phase de "docking" (predictors) où les compétiteurs proposent, à partir des structures non liées des deux partenaires, les 10 amarrages les plus probables selon leur approche
  2. une seconde phase de "scoring" (scorers) où les participants de la phase précédente ont fourni aux organisateurs un ensemble de 100 poses qu'ils considèrent comme les plus vraisemblables selon leur approche. L'ensemble des poses ainsi collecté est mis à disposition des compétiteurs pour évaluer les conformations indépendamment de la phase d'amarrage. Les compétiteurs sont ensuite évalués sur l'ensemble des 10 poses qu'ils auront sélectionnées comme étant les plus probables dans l'ensemble des poses candidates.

La communauté CAPRI est très active et des compétitions sont organisées régulièrement. De 2001 à 2011, 26 rounds ont été organisés, soit un total de 56 structures de complexes proposés à la communauté pour tester et améliorer leurs outils. La structure actuelle de CAPRI divisée en deux phases est apparue en 2005 avec le round 8 (targets 22 et 23). Les nouvelles règles de CAPRI ont été formalisées en 2007.

Cette modification des règles de CAPRI correspondait à une évolution de la communauté et des outils qu'elle avait développé. En effet, lors du second meeting CAPRI organisé en décembre 2004 à Gaeta (Italie), il est devenu évident que plusieurs équipes travaillant sur la problématique de l'amarrage protéine-protéine avait conçu et mis en œuvre des méthodes capables d'isoler les conformations les plus plausibles parmi un large ensemble de conformations et ce indépendamment de l'algorithme d'amarrage utilisé. Sachant que la première phase de l'amarrage, qui consiste à explorer le plus large ensemble d'appariements possibles entre les deux protéines, est la phase la plus coûteuse en temps de calcul. La capacité à évaluer des conformations sans connaître les détails de leur génération, rend donc possible l'interaction entre les approches les plus efficaces de chacune des deux phases.

1 Critères d'évaluation de CAPRI

Les résultats de CAPRI sont évalués selon des critères très stricts qui sont explicités ci-après :

[9] `http ://www.ebi.ac.uk/msd-srv/capri/`

- Haute qualité : $[fnat \geq 0.5$ et $(I_{RMSD} \leq 1$ ou $L_{RMSD} \leq 1)]$
- Moyenne qualité : $[(fnat \geq 0.3$ et $fnnat < 0.5)$ et $(I_{RMSD} \leq 2.0$ ou $L_{RMSD} \leq 5.0)]$ ou $[fnat > 0.5$ et $(I_{RMSD} > 1.0$ ou $I_{RMSD} > 1.0)]$
- Acceptable : $[(fnat \geq 0.1$ et $fnnat < 0.1)$ et $(I_{RMSD} \leq 4.0$ ou $L_{RMSD} \leq 10.0)]$ ou $[fnat > 0.3$ et $(L_{RMSD} > 5.0$ ou $I_{RMSD} > 2.0)]$
- Incorrect sinon

Où $fnat$ est la fraction de contacts natifs présents dans la prédiction, $fnnat$ est la fraction de contacts de la prédiction qui sont natifs, I_{RMSD} est le RMSD entre l'interface prédite et l'interface native, L_{RMSD} est le RMSD entre le ligand prédit et le ligand natif, les récepteurs étant superposés.

2 Résultats

Dans [BAJP07], nous avons appliqué notre modèle prédictif obtenu par algorithme évolutionnaire sur les conformations prédites par Dock pour les rounds 3 à 6 de CAPRI (10 cibles) puis par Haddock pour ces mêmes rounds mais moins de cibles (5 cibles) car Haddock n'avait pas participé à l'intégralité de ces rounds. Nos résultats montrent que notre modèle prédictif permet de réordonner beaucoup plus efficacement l'ensemble des conformations proposées par l'une ou l'autre des méthodes testées. Le modèle prédictif est totalement indépendant des algorithmes d'amarrage utilisés pour générer les conformations. Ces résultats ont contribué à faire évoluer la compétition CAPRI en ajoutant la phase dédiée au "scoring".

Dans [BAPR11], nous présentons une étude comparable sur les rounds 8 à 18 de CAPRI (9 cibles).

Dans ce travail, nous montrons que la génération de conformations candidates et l'évaluation de la complémentarité de forme par Hex, couplées à l'évaluation des caractéristiques physico-chimiques par les empreintes Voronoï, permettent une prédiction particulièrement efficace de la conformation de complexes protéine-protéine. Nos résultats montrent que pour 7 des 9 cibles, notre approche permet d'identifier une conformation de qualité moyenne. Nous avons comparé nos résultats avec ceux de 12 autres approches participant régulièrement à la compétition CAPRI. Parmi ces 12 approches, seules 3 approches (y compris notre précédente approche VDOCK) avait participé à la phase de "scoring" des 9 cibles. Aucune approche n'a été en mesure d'identifier une solution acceptable pour les 2 cibles où nous sommes également en échec. Enfin, parmi tous les résultats obtenus, la combinaison de Hex avec la modélisation de Voronoï permet d'obtenir les meilleurs performances.

D Conclusion

La méthode que nous avons conçue et mis en œuvre après 8 ans de travail est maintenant suffisamment opérationnelle, aussi bien en terme de temps de calcul qu'en terme de précision pour envisager son utilisation dans le cadre de réelles applications biologiques.

Les résultats obtenus par notre méthode sur des complexes relativement peu sujet à déformation lors de l'interaction des deux protéines sont du même niveau que ceux fournis par les meilleures approches recensées dans le cadre de la compétition CAPRI. Bien que plutôt orientée vers la prédiction de complexes protéine-protéine relativement peu sujets à déformation, notre approche permet malgré tout de prédire des complexes dont les protéines se déforment à l'interaction.

Nous avons choisi de faire porter nos efforts sur deux aspects : (i) une modélisation gros-grain des complexes permettant de prendre en compte la déformation des protéines due à l'amarrage. Cette modélisation est réalisée par des diagrammes de Voronoï. Et (ii), l'utilisation de modèles prédictifs appris

à partir de données connues. Ces modèles sont obtenus par une approche d'apprentissage supervisé de type ensemble learning.

Nous avons également choisi d'exploiter les données générées par l'un des meilleurs algorithmes actuellement disponible pour générer des conformations candidates : Hex. L'utilisation par Hex de la transformation de Fourier Rapide sur GPU permet d'obtenir des temps de calculs pour la phase de génération des conformations candidates très faibles. Cette combinaison (Hex + Voronoï + modèle prédictif) nous permet d'obtenir une approche très rapide et dont les prédictions sont comparables avec les meilleurs approches disponibles (en intégrant une contrainte de temps de calcul similaire).

Ces résultats nous permettent d'envisager de nombreuses perspectives de recherches que nous détaillerons dans le chapitre suivant.

Sommaire

A Ré-annotation fonctionnelle de génomes et annotation de mutants
B Cross-docking
C Réseau d'interaction protéine-protéine
D Interactions protéine-ARN

CHAPITRE 4

CONCLUSION, PERSPECTIVES

LES travaux présentés dans ce manuscrit sont orientés vers l'analyse des interactions entre protéines et reflètent en cela une large partie de mes travaux de recherche réalisés depuis 2005.

Si nous replaçons les résultats obtenus sur un axe temporel, nous pouvons constater que j'ai tout d'abord commencé par étudier les interactions physiques en trois dimensions entre les protéines (amarrage protéine-protéine) avant de m'intéresser à la prédiction des interactions entre protéines directement à partir de leur séquences (indépendamment de leurs structures).

Une partie de mes travaux, que j'ai choisi de ne pas présenter ici, concerne la prédiction des fonctions associées aux protéines. Ces travaux ont été essentiellement réalisés dans le cadre de la thèse de Lucie Gentils (2004-2008), thèse encadrée par Christine Froidevaux à laquelle j'ai activement participé à partir de janvier 2006. L'objectif de cette thèse était de proposer un module d'annotation semi-automatique de génome à l'aide d'un vocabulaire contrôlé exprimé sous la forme d'une hiérarchie fonctionnelle [GATN+08, AGTN+08, GAB+06].

Par manque de résultats récents sur le sujet, j'ai choisi de ne pas consacrer de chapitre à ces travaux et aux résultats associés dans ce document. Mais, les résultats obtenus, aussi bien dans la prédiction des interactions protéine-protéine directement à partir des séquences permettent d'envisager sous un nouvel angle mes travaux antérieurs sur la problématique de l'annotation fonctionnelle.

L'annotation fonctionnelle fait ainsi partie des perspectives de recherches à court et moyen terme que j'envisage pour ces travaux.

A Ré-annotation fonctionnelle de génomes et annotation de mutants

Considérons que nous nous trouvons dans un cadre de travail tel qu'un génome de référence est connu et bien annoté ("core genome"). Plusieurs mutants de ce génome sont réalisés afin de modifier certaines propriétés du génome de référence (applications fréquentes dans le domaine de l'agro-alimentaire par exemple). Nous nous situons alors dans un cadre *a priori* idéal pour pouvoir non seulement prédire correctement les interactions entre les protéines de chacun des génomes mutés, mais également pour pouvoir prédire les fonctions associées aux protéines présentes dans le génome muté à partir des fonctions déjà associées aux protéines du génome de référence.

Considérons maintenant que nous nous plaçons dans un cadre plus restrictif que le précédent, à savoir analyse d'un génome déjà séquencé et annoté mais en exploitant des informations nouvelles (nouveaux orthologues par exemple). Dans ce cadre, nos travaux concernant la prédiction des interactions protéine-protéine pourraient être étendus pour intégrer le fait qu'une partie des connaissances est déjà présente (interactions déjà connues entre des protéines) et que ces connaissances pourraient être utilisées pour "calibrer" les différents paramètres intervenant dans la prédictions des interactions (classifieurs, seuils d'élagages, *etc.*).

Pour ces deux cadres de travail, les résultats précédemment obtenus via nos travaux antérieurs concernant l'annotation fonctionnelle, notamment en utilisant des métriques proposées par [KMNF06] pour évaluer nos modèles dans le cadre de vocabulaire défini dans une hiérarchie fonctionnelle, pourront être étendus. Une extension naturelle étant liée à la prise en considération de nouvelles informations fournies sous la forme de probabilités d'interactions entre protéines.

Compte tenu du fait que de la quantité de données disponible augmente très rapidement aujourd'hui, notamment grâce aux nouvelles techniques de séquençage, il est raisonnable de ré-analyser régulièrement les génomes déjà annotés pour vérifier si de nouvelles connaissances peuvent en être extraites.

L'analyse de ces génomes ne se limite pas à la prédiction des protéines en interaction, ni à leur annotation fonctionnelle. Il faut également intégrer la notion de prédiction de site d'interaction et d'interaction physique entre les protéines afin de pouvoir utiliser ces connaissances pour concevoir de nouveaux peptides (drug design) ou pour mieux comprendre les mécanismes régulant le fonctionnement des organismes étudiés.

L'une des perspectives de recherche que je propose s'inscrit dans le contexte de "cross-docking" ou amarrage croisé.

B Cross-docking

Le principe du cross-docking est de prédire l'ensemble des interactions physiques pour toutes les protéines d'un génome. La première étape consiste donc à prédire les interactions protéine-protéine dans le génome, puis, pour les interactions les plus probables à réaliser l'étape de docking (amarrage). Sachant que l'étape de docking est relativement coûteuse en temps et qu'il est nécessaire, contrairement à l'étape de prédiction des interactions protéine-protéine, de disposer d'une structure 3D de chacun des partenaires, il est crucial que la phase de prédiction des interactions protéine-protéine ne produisent pas un nombre trop important de faux positifs.

Parmi les nombreuses pistes de recherche associées au cross-docking, je propose d'en explorer deux. Premièrement, je propose de nous placer dans un cadre d'apprentissage actif. L'objectif est de sélectionner la paire de protéines dont le degré de confiance dans la prédiction d'interaction est le plus élevé, de réaliser l'étape de docking, puis d'exploiter les résultats obtenus par le docking pour mettre à jour le modèle de prédiction des interactions protéine-protéine. En effet, si le score associé à la meilleure conformation 3D obtenue par l'algorithme de Docking utilisé est inférieur à un seuil prédéfini (obtenu en calibrant l'algorithme sur des benchmarks connus), alors la probabilité que la paire de protéines soit en interaction est faible. Nous disposons alors d'un nouveau contre-exemple pour mettre à jour le modèle. Inversement, si le score obtenu par l'algorithme de Docking est supérieur au seuil prédéfini, alors nous disposons d'un nouvel exemple positif d'interaction.

La deuxième piste de recherche que je propose concerne l'utilisation de la Trace Evolutionnaire (ET, Evolutionary Trace) proposée par O. Lichtarge [LW10, LBC96]. Ces travaux montrent qu'il est possible

d'associer, à chaque acide aminé d'une protéine, un score reflétant la probabilité d'être impliqué dans l'interaction avec un partenaire. Cette évaluation ne requiert que la séquence des protéines et elle exploite la phylogénie des orthologues pour inférer quels acides aminés sont conservés ou non dans les différentes espèces considérées. L'utilisation de la trace évolutionnaire nous permettrait soit de filtrer *a priori* les paires de protéines prédites en interaction avant de lancer la phase de docking, soit d'initialiser la phase de docking en se focalisant prioritairement sur les acides aminés susceptibles d'être en interaction. Cette perspective devrait permettre de réduire le nombre de docking à effectuer et ainsi rendre crédible l'utilisation d'une étape de cross-docking dans un contexte haut-débit.

Nous pouvons également envisager nos travaux sous l'angle de la reconstruction complète du réseau d'interaction protéine-protéine.

C Réseau d'interaction protéine-protéine

L'un des objectifs de l'analyse de génomes est de pouvoir mieux comprendre les interactions entre protéines et notamment de pouvoir prédire le réseau d'interaction protéine-protéine le plus complet possible. La connaissance de ce réseau permet de comprendre les différentes voies de régulation des mécanismes complexes d'un organisme et ainsi de pouvoir agir sur quelques éléments clés du réseau afin d'activer ou d'inhiber certaines réactions spécifiques.

Pouvoir prédire des interactions entre protéines de manière fiable représente l'un des éléments permettant d'envisager la prédiction de réseaux d'interaction.

Je pense qu'une piste de recherche intéressante dans ce contexte serait de formaliser les réseaux d'interactions sous la forme de graphes formels et ainsi pouvoir utiliser les méthodes associées telles que l'analyse de concepts formels. Jusqu'à récemment seuls des graphes impliquant des relations binaires entre concepts étaient analysables par les méthodes d'analyse formelle. [Kay11] a étendu le cadre classique en y ajoutant la possibilité de traiter des relations valuées (à valeur réelle) entre concepts.

Le graphe se représente sous la forme d'une matrice protéine$\times$protéines (matrice triangulaire supérieure) où la confiance dans l'interaction entre deux protéines est matérialisée par une valeur réelle située dans l'intervalle $[0, 1]$ où 0 représente la certitude qu'il n'existe pas de relation entre les deux protéines et 1 la certitude que les deux protéines sont en interaction (interaction vérifiée expérimentalement par exemple). Les méthodes d'analyse formelle appliquées à un tel graphe nous permettrons d'identifier des motifs qui pourront correspondre soit à des complexes, soit à des voies de régulations. L'utilisation d'autres réseaux connus (orthologues par exemple) nous permettra de valider les connaissances trouvées.

Enfin, depuis septembre 2011, je me suis engagé sur une nouvelle voie : l'étude des interactions protéine-ARN.

D Interactions protéine-ARN

Je travaille actuellement, dans le cadre de la thèse d'Adrien Guilhot-Gaudeffroy débutée en septembre 2011, à l'étude de l'amarrage protéine-ARN. Cette thèse est co-encadrée par Christine Froidevaux, Julie Bernauer et moi-même.

Les résultats obtenus dans le cadre protéine-protéine sont en partie réexploités. Le fait de remplacer une protéine par un ARN, nous impose de redéfinir plusieurs éléments dont la génération des candidats, la modélisation de Voronoï et les fonctions d'évaluation. Les résultats préliminaires associés à ce travail

de thèse ont été présentés sous la forme d'un poster [GGAB11].

BIBLIOGRAPHIE

[AAA+05] E. ALPHONSE, A. AMRANI, J. AZÉ, T. HEITZ, A.-D. MEZAOUR, et M. ROCHE. « Préparation des données et analyse des résultats de DEFT'05 ». Dans *Dans les actes de la conférence "Traitement Automatique des Langues Naturelles" (TALN 2005) - Atelier DEFT'05*, volume 2, pages 99–111, 6-10 juin 2005. 1

[AAH+04] A. AMRANI, J. AZÉ, T. HEITZ, Y. KODRATOFF, et M. ROCHE. « From the texts to the concepts they contain : a chain of linguistic treatments ». Dans *In Proceedings of TREC'04 (Text REtrieval Conference), National Institute of Standards and Technology, Gaithersburg Maryland USA, 2004*, pages 712–722, 2004. 1

[ABH+11] J. AZÉ, T. BOURQUARD, S. HAMEL, A. POUPON, et D.W. RITCHIE. « Using Kendall-Tau Meta-Bagging to Improve Protein-Protein Docking Predictions ». Dans M. Loog et AL., éditeur, *PRIB 2011*, pages 284–295, 2011. . . . . 31, 35

[AGTN+08] J. AZÉ, L. GENTILS, C. TOFFANO-NIOCHE, J.-F. LOUX, V. Gibrat, P. BESSIÈRES, A. ROUVEIROL, C. Poupon, et C. FROIDEVAUX. « Towards a semi-automatic functional annotation tool based on decision-tree techniques ». *BMC Proceedings 2008*, 2((Suppl 4) :S3), 2008. 39

[AIS93] R AGRAWAL, Tomasz IMIELINSKI, et Arun SWAMI. « Mining association rules between sets of items in large databases ». *ACM SIGMOD Record*, 22(2) :207–216, Jan 1993. 16, 17

[AK02] J. AZÉ et Y. KODRATOFF. « A study of the Effect of Noisy Data in Rule Extraction Systems ». Dans *Proceedings of the Sixteenth European Meeting on Cybernetics and Systems Research (EMCSR'02)*, volume 2, pages 781–786, 2002. 24

[ALLV07] J. AZÉ, P. LENCA, S. LALLICH, et B. VAILLANT. « A study of the robustness of association rules ». Dans R. STAHLBOCK, S. F. CRONE, et CSREA Press S. LESSMANN, éditeurs, *The 2007 International Conference on Data Mining (DMIN'07)*, pages 132–137, 2007. 5

[ALS03] J. AZÉ, N. LUCAS, et M. SEBAG. « A New Medical Test for Atherosclerosis Detection : GeNo », september 2003. . . . . 1, 11

[ALS04] J. AZÉ, N. LUCAS, et M. SEBAG. « A Genetic ROC-based Classifier », July 2004. . 1, 5, 10

[ARKS04] J. AZÉ, M. ROCHE, Y. KODRATOFF, et M. SEBAG. « Learning to Order Terms : Supervised Interestingness Measures in Terminology Extraction ». Dans *International Conference on Computational Intelligence*, pages 478–481, 2004. 11

[ARKS05] J. AZÉ, M. ROCHE, Y. KODRATOFF, et M. SEBAG. « Preference Learning in Terminology Extraction : A ROC-based approach ». Dans *In proceedings of ASMDA'05 (Applied Stochastic Models and Data Analysis), may 2005, Brest, France*, pages 209–219, 2005. 1, 5, 11

[ARS05a] J. AZÉ, M. ROCHE, et M. SEBAG. « Bagging Evolutionary ROC-based Hypotheses Application to Terminology Extraction ». Dans *In proceedings of ROCML (ROC Analysis in Machine Learning), Bonn, Germany*, 2005. 1

[ARS05b] Jérôme AZÉ, Mathieu ROCHE, et Michèle SEBAG. « Bagging Evolutionary ROC-based Hypotheses Application to Terminology Extraction. ». *ICML 2005 workshop on ROC Analysis in Machine Learning*, Jan 2005. 11

[AS94a] R AGRAWAL et R SRIKANT. « Fast algorithms for mining association rules ». *Proc. 20th Int. Conf. Very Large Data Bases*, pages 487–499, Jan 1994. . . . . 16, 17

[AS94b] R. AGRAWAL et R. SRIKANT. « Fast Algorithms for Mining Association Rules ». Dans Jorge B. BOCCA, Matthias JARKE, et Carlo ZANIOLO, éditeurs, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994. 24

[AS01] H ABBASS et R SARKER. . . . « PDE : A Pareto-frontier differential evolution approach for multi-objective optimization problems ». . . . , Jan 2001. 9

[AZ12] Charu C AGGARWAL et Chengxiang ZHAI. « A Survey of Text Clustering Algorithms ». *Mining Text Data*, pages 77–128, 2012. 16

[Azé03] J. AZÉ. « Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances ». Dans *Revue RIA-ECA numéro spécial EGC03*, volume 17 (1,2,3), pages 171–182, 2003. 5, 17

[BAJP05] J. BERNAUER, J. AZÉ, J. JANIN, et A. POUPON. « Une nouvelle fonction de score pour l'amarrage protéine-protéine fondée sur les diagrammes de Voronoï ». Dans *In proceedings of Journées Ouvertes de Biologie Informatique Mathématiques (JOBIM)*, pages 459–469, 2005. . . . . 11, 34

[BAJP07] J. BERNAUER, J. AZÉ, J. JANIN, et A. POUPON. « A new protein-protein docking scoring function based on interface residue properties. ». *Bioinformatics*, 5(23) :555–62, 2007. . . . . 11, 31, 34, 35, 36

[BAPR11] T. BOURQUARD, J. AZÉ, A. POUPON, et D.W. RITCHIE. « Amarrage protéine-protéine par couplage de la complémentarité de forme et des empreintes Voronoï ». Dans *Jobim 2011*, pages 9–16, 2011. . . . . 31, 36

[BBAP09] T. BOURQUARD, J. BERNAUER, J. AZE, et A. POUPON. « Comparing Voronoi and Laguerre tessellations in the protein-protein docking context ». Dans *Sixth International Symposium on Voronoi Diagrams (ISVD)*, pages 225–232, 2009. . . . . 31, 34

[BBAP11] T. BOURQUARD, J. BERNAUER, J. AZÉ, et A. POUPON. « A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions ». *PLoS One*, 6(4) :e18541, 04 2011. . . . . 3, 15, 31, 34, 35

[BDP+02] Jean-Daniel BOISSONNAT, Olivier DEVILLERS, Sylvain PION, Monique TEILLAUD, et Mariette YVINEC. « Triangulations in CGAL ». *Computational Geometry : Theory & Applications*, 22 :5–19, Aug 2002. 34

[Ber02] Pavel BERKHIN. « Survey of Clustering Data Mining Techniques ». *Technical Report*, pages 1–56, Jun 2002. 16

[Ber06] Julie BERNAUER. « Utilisation de la tessellation de Voronoï pour l'étude des complexes protéine-protéine ». *PhD*, pages 1–166, Jun 2006. . . . . 33, 34

[BK98] Eric BAUER et Ron KOHAVI. « An Empirical Comparison of Voting Classification Algorithms : Bagging, Boosting and Variants ». *Machine Learning*, pages 1–38, Sep 1998. 11

[BMS97] S. BRIN, R. MOTWANI, et C. SILVERSTEIN. « Beyond market baskets : generalizing association rules to correlations ». Dans *Proceedings of ACM SIGMOD'97*, pages 265–276, 1997. . . . . 17, 24

[Bor10] Christian BORGELT. « Simple algorithms for frequent item set mining ». *Advances in Machine Learning II*, Jan 2010. 16

[Bou09] Thomas BOURQUARD. « Exploitation des algorithmes génétiques pour la prédiction de structures protéine-protéine ». *PhD*, pages 1–133, Nov 2009. 34

[BPAJ05] J. BERNAUER, A. POUPON, J. AZÉ, et J. JANIN. « A docking analysis of the statistical physics of protein-protein recognition ». *Phys Biol.*, 2(2) :S17–23, 2005. . . . . 11, 31, 34

[Bre96] Leo BREIMAN. « Bagging predictors ». *Machine Learning*, 24(2) :123–140, 1996. . . 11

[Buc62] J BUCHANAN. « The relevance of Pareto optimality ». *Journal of conflict resolution*, Jan 1962. 9

[BWF+00] Helen M BERMAN, John WESTBROOK, Zukan FENG, Gary GILLILAND, T.N BHAT, Helge WEISSIG, Ilya N SHINDYALOV, et Philip E BOURNE. « The protein data bank ». *Nucleic Acids Research*, 28(1) :235–242, Jan 2000. 33

[CDJ91] Jacqueline CHERFILS, Stéphane DUQUERROY, et Joël JANIN. « Protein-protein recognition analyzed by docking simulation ». *Proteins*, 11(4) :271–280, Jan 1991. 33

[CH90] K. W. CHURCH et P. HANKS. « Word Association Norms, Mutual Information, and Lexicography ». *Computational Linguistics*, 16 :22–29, 1990. 24

[CL07] Roger A CRAIG et Li LIAO. « Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices ». *BMC Bioinformatics*, 8(6), 2007. . . . . 25, 26

[CM02] A. CORNUÉJOLS et L. MICLET. *Apprentissage Artificiel - Concepts et algorithmes*. Cépadues, 2002. 6

[Coh60] Jacob COHEN. « A coefficient of agreement for nominal scales ». *Educational and psychological measurement*, pages 37–47, Jan 1960. 24

[DBB03] Cyril DOMINGUEZ, Rolf BOELENS, et Alexandre M J J BONVIN. « HADDOCK : a protein-protein docking approach based on biochemical or biophysical information ». *Journal of the American Chemical Society*, 125 :1731–1737, Jan 2003. . . . . 3, 4

[Deb03] K DEB. « Multi-objective evolutionary algorithms : Introducing bias among Pareto-optimal solutions ». *Advances in evolutionary computing*, Jan 2003. 9

[Dic45] Lee R. DICE. « Measures of the Amount of Ecologic Association Between Species ». *Ecology*, 26(3) :297–302, 1945. 24

[DMS+05] Michael D DAILY, David MASICA, Arvind SIVAUBRAMANIAN, Sony SOMAROUTHU, et Jeffrey J GRAY. « CAPRI rounds 3–5 reveal promising successes and future challenges for RosettaDock ». *Proteins*, 60 :181–186, Jan 2005. 33

[DSHB98] T DANDEKAR, B SNEL, M HUYNEN, et P BORK. « Conservation of gene order : a fingerprint of proteins that physically interact ». *Trends in Biochemical Sciences*, 23(9) :324–328, SEP 1998. 20

[dVDA12] de VIENNE D.M. et J. AZÉ. « Efficient Prediction of Co-Complexed Proteins Based on Coevolution ». *PLoS One*, page under submission, 2012. . . . . 15, 19, 21, 25, 26

[dVGM07] Damien M. de VIENNE, Tatiana GIRAUD, et Olivier C. MARTIN. « A congruence index for testing topological similarity between trees ». *Bioinformatics*, 23(23) :3119–3124, 2007. 23

[FF03] Johannes FÜRNKRANZ et Peter A FLACH. « An analysis of rule evaluation metrics ». *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Jan 2003. 8

[FFHO02] C FERRI, P FLACH, et J HERNÁNDEZ-ORALLO. « Learning decision trees using the area under the ROC curve ». *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 139–146, 2002. 10

[FkS89] Stanley FIELDS et Ok kyu SONG. « A novel genetic system to detect protein protein interactions ». *Letters to Nature*, 340 :245–246, Jan 1989. 20

[FS97] Yoav FREUND et Robert E SCHAPIRE. « A decision-theoretic generalization of on-line learning and an application to boosting ». *Journal of Computer and System Sciences*, 55 :119–139, Jan 1997. 12

[GAB+06] L. GENTILS, J. AZÉ, P. BESSIÈRES, J.-F. GIBRAT, V. LOUX, C. ROUVEIROL, et C. FROIDEVAUX. « Learning Rules for Microbial Genome Annotation ». Dans *Automated Function Prediction, 2006*, pages 82–83, San Diego, CA, USA, 2006. poster, abstract. 39

[Gan87] Jean-Gabriel GANASCIA. « CHARADE : A Rule System Learning System ». Dans *IJCAI*, pages 345–347, 1987. 24

[GATN+08] L. GENTILS, J. AZÉ, C. TOFFANO-NIOCHE, V. LOUX, A. POUPON, J.-F. GIBRAT, et C. FROIDEVAUX. « Mesures Hiérarchiques pondérées pour l'évaluation d' un système semi-automatique d'annotation de génomes utilisant des arbres de décision ». Dans *Extraction et gestion des connaissances (EGC'2008)*, volume 1, pages 133–138, 2008. . 39

[GBJ+00] CS GOH, AA BOGAN, M JOACHIMIAK, D WALTHER, et FE COHEN. « Co-evolution of proteins with their interaction partner ». *J Mol Biol*, 299 :283–293, 2000. 23

[GGAB11] Adrien GUILHOT-GAUDEFFROY, Jérôme AZÉ, et Julie BERNAUER. « Prediction of protein-nucleic acid interactions : building a scoring function with the Voronoi diagram. », June 2011. 42

[GJJE+10] Beatriz GARCÍA-JIMÉNEZ, David JUAN, Iakes EZKURDIA, Eduardo ANDRÉS-LEÓN, et Alfonso VALENCIA. « Inference of Functional Relations in Predicted Protein Networks with a Machine Learning Approach ». *PLoS ONE*, 5(4), 2010. 21

[GMCF10] Srinivasa M GOPAL, Shayantani MUKHERJEE, Yi-Ming CHENG, et Michael FEIG. « PRIMO/PRIMONA : A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy ». *Proteins*, 78 :1266–1281, Jan 2010. 33

[GMW+03] J.J GRAY, S MOUGHON, C WANG, O SCHUELER-FURMAN, B KUHLMAN, C.A ROHL, et D BAKER. « Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations ». *Journal of molecular biology*, 331(1) :281–299, 2003. 33

[Goo03] Irving John GOOD. *The Estimation of Probabilities : An Essay on Modern Bayesian Methods*. The MIT Press Classics Series, 2003. 24

[Gra06] Jeffrey J GRAY. « High-resolution protein–protein docking ». *Current Opinion in Structural Biology*, 16 :183–193, Jan 2006. 33

[GSST07] C GAGNÉ, M SEBAG, M SCHOENAUER, et M TOMASSINI. « Ensemble learning for free with evolutionary algorithms ? ». *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1782–1789, 2007. 11

[Han90] L HANSEN. . . . « Neural network ensembles ». *Pattern Analysis and Machine . . .*, Jan 1990. 11

[HFH+09] M. HALL, E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, et I. H. WITTEN. « The WEKA Data Mining Software : An Update ». *SIGKDD Explorations*, 11(1) :10–18, 2009. 23

[ITM+00] Takashi ITO, Kosuke TASHIRO, Shigeru MUTA, Ritsuko OZAWA, Tomoko CHIBA, Mayumi NISHIZAWA, Kiyoshi YAMAMOTO, Satoru KUHARA, et Yashiyuki SAKAKI. « Toward a protein–protein interaction map of the budding yeast : a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins ». *Proceedings of the National Academy of Sciences*, 97(3) :1143–1147, Jan 2000. . . 20

[Jac08] P. JACCARD. « Nouvelles recherches sur la distribution florale ». *Bulletin de la Société Vaudoise en Sciences Naturelles*, 44 :223–270, 1908. 24

[JMF99] A.K JAIN, M.N MURTY, et P.J FLYNN. « Data clustering : a review ». *ACM computing surveys (CSUR)*, 31(3) :264–323, 1999. 16

[JPV08] David JUAN, Florencio PAZOS, et Alfonso VALENCIA. « High-confidence prediction of global interactomes based on genome-wide coevolutionary networks ». *PNAS*, 105(3) :934–939, 2008. . . . . 23, 25, 26

[JYG+03] R JANSEN, HY YU, D GREENBAUM, Y KLUGER, NJ KROGAN, SB CHUNG, A EMILI, M SNYDER, JF GREENBLATT, et M GERSTEIN. « A Bayesian networks approach for predicting protein-protein interactions from genomic data ». *Science*, 302(5644) :449–453, OCT 17 2003. 21

[KARMT03] Y. KODRATOFF, J. AZÉ, M. ROCHE, et O. MATTE-TAILLIEZ. « Des textes aux associations entre les concepts qu'ils contiennent ». *Dans les actes des XXXVièmes Journées de Statistique (résumé) Volume 2, p599-602 et version complète dans le numéro spécial de la revue RNTI (Revue des Nouvelles Technologies de l'Information) "Entreposage et Fouille de données"*, 1 :171–182, 2003. 1

[Kay11] Mehdi KAYTOUE. « Traitement de données numériques par analyse formelle de concepts et structures de patrons ». *PhD*, pages 1–135, May 2011. 41

[KMNF06] Svetlana KIRITCHENKO, Stan MATWIN, Richard NOCK, et A. Fazel FAMILI. « Learning and Evaluation in the Presence of Class Hierarchies : Application to Text Categorization ». Dans Luc LAMONTAGNE et Mario MARCHAND, éditeurs, *Canadian Conference on AI*, volume 4013 de *Lecture Notes in Computer Science*, pages 395–406. Springer, 2006. 40

[KV95] Anders KROGH et Jesper VEDELSBY. « Neural network ensembles, cross validation, and active learning ». *Advances in neural information processing systems*, 7 :231–238, Jan 1995. 11

[LA07] I.-C. LERMAN et J. AZÉ. « *A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link* », pages 207–236. Springer, 2007. 5

[LAS02] N. LUCAS, J. AZÉ, et M. SEBAG. « Atherosclerosis Risk Identification and Visual Analysis », september 2002. 1

[LBC96] O. LICHARTGE, H.R. BOURNE, et F.E. COHEN. « An evolutionary trace method defines binding surfaces common to protein families ». *Journal of Molecular Biology*, 257(2) :342–358, 1996. 40

[LEG+09] Eduardo Andres LEON, Iakes EZKURDIA1, Beatriz GARCIA, Alfonso VALENCIA, et David JUAN. « EcID. A database for the inference of functional interactions in E. coli ». *Nucleic Acids Research*, 37(Database issue) :D629–D635, 2009. 21

[LFZ99] N. LAVRAC, P. FLACH, et B. ZUPAN. « Rule Evaluation Measures : A Unifying View ». Dans S. DŽEROSKI et P. FLACH, éditeurs, *Ninth International Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 de *Lecture Notes in Artificial Intelligence*, pages 174–185. Springer-Verlag, June 1999. . . . . 17, 24

[LG08] S LYSKOV et J. J GRAY. « The RosettaDock server for local protein-protein docking ». *Nucleic Acids Research*, 36(Web Server) :W233–W238, May 2008. . . . . 3, 4, 33

[LHZ03a] Charles X LING, Jin HUANG, et Harry ZHANG. « AUC : a better measure than accuracy in comparing learning algorithms ». *Advances in Artificial Intelligence*, Jan 2003. . . . 10

[LHZ03b] Charles X LING, Jin HUANG, et Harry ZHANG. « AUC : a statistically consistent and more discriminating measure than accuracy ». *International joint Conference on artificial intelligence*, pages 519–524, Jan 2003. . . . . 8, 10

[LMP+03] P LENCA, P MEYER, P PICOUET, B VAILLANT, et S LALLICH. « Critères d'évaluation des mesures de qualité des règles d'association ». *Revue des Nouvelles Technologies de l'information (RNTI) RNTI*, 1 :123–134, 2003. 17

[Loe47] J. LOEVINGER. « A systematic approach to the construction and evaluation of tests of ability ». *Psychological Monographs*, 61 :1–49, 1947. 24

[LR10] Simon C. LOVELL et David L. ROBERTSON. « An Integrated View of Molecular Coevolution in Protein Protein Interactions ». *Molecular Biology and Evolution*, 27(11) :2567–2575, 2010. 20

[LT04a] S. LALLICH et O. TEYTAUD. « Évaluation et validation de l'intérêt des règles d'association », avril 2004. 17

[LT04b] S. LALLICH et O. TEYTAUD. « Évaluation et validation de l'intérêt des règles d'association », avril 2004. 24

[LV11] Shiyong LIU et Ilya A VAKSER. « DECK : Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking ». *BMC Bioinformatics*, 12(280) :1–7, Jan 2011. 33

[LVL07] Stéphane LALLICH, Benoît VAILLANT, et Philippe LENCA. « A probabilistic framework towards the parameterization of association rule interestingness measures ». . . . *and Computing in Applied Probability*, 9 :447–463, Jan 2007. 17

[LW10] Olivier LICHTARGE et Angela WILKINS. « Evolution : a guide to perturb protein function and networks ». *Current Opinion in Structural Biology*, 20(3) :351–359, Jun 2010. . . 40

[LWJ+04] Nan LIN, Baolin WU, Ronald JANSEN, Mark GERSTEIN, et Hongyu ZHAO. « Information assessment on predicting protein-protein interactions ». *BMC Bioinformatics*, 5(1) :154, 2004. 21

[LXP+05] L. J. LU, Y. XIA, A. PACCANARO, H. YU, et M. GERSTEIN. « Assessing the Limits of Genomic Data Integration for Protein-Protein Interactions ». *Genome Research*, 15 :945–953, 2005. 21

[MAB11] I. H MOAL, R AGIUS, et P. A BATES. « Protein-protein binding affinity prediction on a diverse set of structures ». *Bioinformatics*, 27(21) :3002–3009, Nov 2011. 33

[Met78] Charles E METZ. « Basic principles of ROC analysis ». *Seminars in nuclear medicine*, VIII(4) :283–298, Jan 1978. 9

[MFKT11] John MOULT, Krzysztof FIDELIS, Andriy KRYSHTAFOVYCH, et Anna TAMONTANO. « Critical assessment of methods of protein structure prediction (CASP)—round IX ». *Proteins*, 79(Suppl 10) :1–5, Jan 2011. 3

[MGC+05] Venkatraman MOHAN, Alan C GIBBS, Maxwell D CUMMINGS, Edward P JAEGER, et Renee L DESJARLAIS. « Docking : successes and challenges ». *Current Pharmaceutical Design*, 11 :323–333, Jan 2005. 32

[MKTE01] J.C. MITCHELL, R. KERR, et LF. TEN EYCK. « Rapid atomic density methods for molecular shape characterization ». *J Mol Graph Model*, 19 :325–30, 2001. 34

[ML12] Juliette MARTIN et Richard LAVERY. « Arbitrary protein-protein docking targets biologically relevant interfaces ». *BMC Biophys*, 5(1) :7, Jan 2012. 3

[OM99] David OPITZ et Richard MACLIN. « Popular ensemble methods : An empirical study ». *Journal of Artificial Intelligence Research*, 11 :169–198, Jan 1999. 11

[Pea00] K. PEARSON. « On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling », 1900. 24

[Pea88] J. PEARL. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988. 24

[PMT+99] M PELLEGRINI, EM MARCOTTE, MJ THOMPSON, D EISENBERG, et TO YEATES. « Assiging protein functions by comparative genome analysis : protein phylogenetic profiles ». *PNAS*, 96 :4285–4288, 1999. 20

[Pou04] Anne POUPON. « Voronoi and Voronoi-related tessellations in studies of protein structure and interaction ». *Current Opinion in Structural Biology*, 14 :233–241, Jan 2004. . . . 33

[PRJS05] Florencio PAZOS, Juan A. G. RANEA, David JUAN, et Michael J. E. STERNBERG. « Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome ». *J. Mol. Biol.*, 352 :1002–1015, 2005. . . . . 20, 23, 25, 26

[PV01] Florencio PAZOS et Alfonso VALENCIA. « Similarity of phylogenetic trees as indicator of protein–protein interaction ». *Protein Engineering*, 14(9) :609–614, 2001. 20, 23, 25, 26

[QBJKS06] Y. QI, Z. BAR-JOSEPH, et J. KLEIN-SEETHARAMAN. « Evaluation of different biological data and computational classification methods for use in protein interaction prediction ». *PROTEINS : Structure, Function, and Bioinformatics*, 63(3) :490–500, 2006. 21

[QN08] Jian QIU et William Stafford NOBLE. « Predicting Co-Complexed Protein Pairs from Heterogeneous Data ». *PLoS Comput Biol*, 4(4) :e1000054, 04 2008. 21

[Qui96] J Ross QUINLAN. « Bagging, boosting, and C4. 5 ». *Proceedings of the National Conference on Artificial Intelligence*, pages 725–730, Jan 1996. 11

[Rak04] A RAKOTOMAMONJY. « Optimizing area under ROC curves with SVMs ». *ROCAI'04*, Jan 2004. 10

[RAKS04] M. ROCHE, J. AZÉ, Y. KODRATOFF, et M. SEBAG. « Learning Interestingness Measures in Terminology Extraction. A ROC-based approach ». Dans *Proceedings of "ROC Analysis in AI" Workshop (ECAI 2004), 22 Août 2004, Valencia, Espagne*, pages 81–88, 2004. . 1

[RE10] D.V.S RAVIKANT et Ron ELBER. « PIE-Efficient filters and coarse grained potentials for unbound protein-protein docking ». *Proteins*, 78 :400–419, Jan 2010. 33

[RHJ+12] Jüri REIMAND, Shirley HUI, Shobhit JAIN, Brian LAW, et Gary D BADER. « Domain-mediated protein interaction prediction : From genome to network ». *FEBS Letters*, pages 1–13, May 2012. 3

[Rit08] D.W. RITCHIE. « Recent progress and future directions in protein-protein docking. ». *Curr Protein Pept Sci*, 9(1) :1–15, 2008. 32

[RK00] David W RITCHIE et Graham J.L KEMP. « Protein docking using spherical polar Fourier correlations ». *PROTEINS : Structure, Function and Genetics*, 39 :178–194, Jan 2000. 3, 34

[RV10] D.W. RITCHIE et V. VENKATRAMAN. « Ultra-fast FFT protein docking on graphics processors ». *Bioinformatics*, 26(19) :2398–405, 2010. 34

[SAL03] M. SEBAG, J. AZÉ, et N. LUCAS. « Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning ». Dans *Proceedings of the Third IEEE International Conference on Data Mining, ICDM 2003*, 2003. . . . . 1, 11

[Sch99] R SCHAPIRE. « A brief introduction to boosting ». *International joint Conference on artificial intelligence*, Jan 1999. 12

[SLA03] M. SEBAG, N. LUCAS, et J. AZÉ. « ROC-based Evolutionary Learning : Application to Medical Data Mining ». Dans *Proceedings of the 6th International Conference on Artificial Evolution, EA 2003*, 2003. . . . . 5, 11

[SYKT05] Tetsuya SATO, Yoshihiro YAMANISHI, Minoru KANEHISA, et Hiroyuki TOH. « The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships ». *Bioinformatics*, 21 :3482–3489, 2005. . . . . 25, 26

[Toz05] Valentina TOZZINI. « Coarse-grained models for proteins ». *Current Opinion in Structural Biology*, 15 :144–150, Jan 2005. 33

[VC06] Miha VUK et Tomaz CURK. « Roc curve, lift chart and calibration plot ». *Metodoloski zvezki*, 3(1) :89–108, Jan 2006. 9

[VCB11] Marc VIDAL, Michael E CUSICK, et Albert-László BARABÁSI. « Interactome networks and human disease ». *Cell*, 144(6) :986–998, Jan 2011. 3

[WKQ+08] X WU, V KUMAR, J Ross QUINLAN, J GHOSH, Q YANG, H MOTODA, G.J MCLACHLAN, A NG, B LIU, et P.S YU. « Top 10 algorithms in data mining ». *Knowledge and Information Systems*, 14(1) :1–37, 2008. 16

[Wol92] David H WOLPERT. « Stacked Generalization ». *Neural Networks*, 5(2) :1–57, Mar 1992. 11, 12

[WS11] Chun WU et Joan-Emma SHEA. « Coarse-grained models for protein aggregation ». *Current Opinion in Structural Biology*, 21 :209–220, Jan 2011. 33

[WSFB05] Chu WANG, Ora SCHUELER-FURMAN, et David BAKER. « Improved side-chain modeling for protein-protein docking ». *Protein Science*, 14 :1328–1339, Jan 2005. 33

[WvD09] Thomas R WEIKL et Carola von DEUSTER. « Selected-fit versus induced-fit protein binding : Kinetic differences and mutational analysis ». *Proteins*, 75(1) :104–110, Jan 2009. 3

[YAR11] Elizabeth YURIEV, Mark AGOSTINO, et Paul A RAMSLAND. « Challenges and advances in computational docking : 2009 in review ». *Journal Of Molecular Recognition*, 24 :149–164, Jan 2011. 32

[Zha00] T ZHANG. « Association rules ». *PAKDD 2000*, LANI 1805 :245–256, Jan 2000. . . 24

Partie I

Curriculum vitae

Sommaire

Situation Professionnelle
Informations Personnelles
Domaine de recherche

Notice Individuelle

Situation Professionnelle

Jérôme Azé
Maître de Conférences Orsay Polytech'Paris-Sud,
Laboratoire de Recherche en Informatique,
Équipe bioinfo-AMIB,
Université Paris-Sud, Bât. 650,
91405 Orsay Cedex.

Informations Personnelles

39, rue des chênes,
91140 Villebon-sur-Yvette.
Né le 29 août 1973,
Marié, trois enfants.

Mél : `jerome.aze@lri.fr`
Web : `www.lri.fr/~aze`

Domaine de recherche

Fouille de Données, Apprentissage, Bioinformatique, Qualité des données, Algorithmes évolutionnaires, Annotation fonctionnelle de protéines, Docking de protéines.

Sommaire

Situation actuelle
Titres Universitaires
Parcours post-doctoral
Thèse
Parcours de premier et second cycle

Parcours et Formation Universitaire

Situation actuelle

2011-2012	Un semestre de délégation INRIA dans l'équipe-projet AMIB + un demi-service d'enseignement (96h eq TD).
2010–2011	Délégation INRIA à temps plein dans l'équipe-projet AMIB.
2005–	Maître de Conférences à Polytech'Paris-Sud, dans l'équipe Bioinformatique du LRI (Bioinfo, dirigée par Mme Christine Froidevaux).

Titres Universitaires

- **Thèse** de l'Université Paris Sud, soutenue en décembre 2003
- **D.E.A.** d'Informatique de l'Université Rennes 1, Mention Assez Bien, 1999.
- **Maitrise** d'informatique de l'Université Rennes 1, Mention Bien, 1998.
- **Licence** d'informatique de l'Université Rennes 1, Mention Bien, 1997.
- **DEUG MIAS** (Maths-Informatique), Université Rennes 1, Mention Assez Bien, 1996.
- **Bac C**, 1991.

Parcours post-doctoral

2003–2005 1/2 ATER à l'Université Orsay Paris-Sud, équipe I&A.

Thèse

Extraction de Connaissances dans des Données Numériques et Textuelles, J. Azé, Thèse de l'Université d'Orsay Paris Sud, Décembre 2003. Soutenue le 16 décembre 2003 devant le Jury composé de Marie-Christine ROUSSET (Présidente), Israël-César LERMAN (Rapporteur), Jean-Daniel ZUCKER (Rapporteur), Jean-Marc PETIT (Examinateur), et Yves KODRATOFF (Directeur de thèse)

Résumé : Le travail réalisé dans le cadre de cette thèse concerne l'extraction de connaissances dans des données transactionnelles. L'analyse de telles données est souvent contrainte par la définition d'un support minimal utilisé pour filtrer les connaissances non intéressantes. Les experts des données ont souvent des difficultés pour déterminer ce support. Nous avons proposé une méthode permettant de ne pas fixer un support minimal et fondée sur l'utilisation de mesures de qualité. Nous nous sommes focalisés sur l'extraction de connaissances de la forme « règles d'association ». Ces règles doivent vérifier un ou plusieurs critères de qualité pour être considérées comme intéressantes et proposées à l'expert. Nous avons proposé deux mesures de qualité combinant différents critères et permettant d'extraire des règles intéressantes.

Nous avons ainsi pu proposer un algorithme permettant d'extraire ces règles sans utiliser la contrainte du support minimal. Le comportement de notre algorithme a été étudié en présence de données bruitées et nous avons pu mettre en évidence la difficulté d'extraire automatiquement des connaissances fiables à partir de données bruitées. Une des solutions que nous avons proposée consiste à évaluer la résistance au bruit de chaque règle et d'en informer l'expert lors de l'analyse et de la validation des connaissances obtenues.

Enfin, une étude sur des données réelles a été effectuée dans le cadre d'un processus de fouille de textes. Les connaissances recherchées dans ces textes sont des règles d'association entre des concepts définis par l'expert et propres au domaine étudié. Nous avons proposé un outil permettant d'extraire les connaissances et d'assister l'expert lors de la validation de celles-ci. Les différents résultats obtenus montrent qu'il est possible d'obtenir des connaissances intéressantes à partir de données textuelles en minimisant la sollicitation de l'expert dans la phase d'extraction des règles d'association.

Parcours de premier et second cycle

2000–2003 Allocataire de Recherche Moniteur à l'Université Orsay Paris Sud, équipe I&A.

1999–2000 Scientifique du Contingent, École militaire de Paris.

–1999 (Parcours de premier et second cycle) : Bac E, DEUG MIAS (Mention Assez Bien, Rennes 1), Licence Informatique (Mention Bien, Rennes 1), Maîtrise Informatique (Mention Bien, Rennes 1), DEA Informatique (Mention Assez Bien, Rennes 1).

Sommaire

Principaux Enseignements
Production de documents pédagogiques

Activités d'enseignements

Principaux Enseignements

- niveau M1 – M2
  - Master 2 BIBS : Intégration et Analyse de Données Biologiques issues du Web (Cours : 80h eq TD)
  - Master 2 Recherche Université Paris-Sud : Fondements de la Représentation des Connaissances (Cours : 47h eq TD)
  - Master 2 Recherche Telecom Bretagne : Introduction à la bioinformatique, fouille de données et fouille de textes (Cours : 36h eq TD)
  - Master 2 Pro : Web Sémantique (TD : 24h), Extraction de Connaissances (Cours : 42h eq TD, TD : 10h)
  - Master CCI : Analyse et Algorithmique (Cours : 131h eq TD)
  - Polytech'Paris-Sud 4e et 5e, Apprentissage : Résolution de Contraintes (Cours : 8h eq TD, TD : 58h, TP : 14h eq TD)

- niveau L1 – L2 – L3
  - Polytech'Paris-Sud 3e année : Algorithmique, Langage C, Projet C : (Cours : 57h eq TD, TD : 30h,TP : 76h eq TD), Système (Cours : 71h eq TD, TD : 12h, TP : 31h eq TD)
  - Prépa Polytech'Paris-Sud 2e année : Algorithmique et Complexité (Cours : 100h eq TD, TD : 82h eq TD, TP : 83h eq TD)

Production de documents pédagogiques

- Mise en place d'un site web d'aide à la programmation en C
  `http ://www.lri.fr/∼aze/page_c/aide_c/`
- Supports de cours en Algorithmique et Langage C

Sommaire

Responsabilités pédagogiques
Commissions de spécialistes
Autres responsabilités – vie du laboratoire

Activités liées à l'administration

Responsabilités pédagogiques

- de juin 2007 à septembre 2010, directeur des études de la 3eme année de la formation initiale de Polytech'Paris-Sud (anciennement IFIPS).
- de septembre 2005 à septembre 2009, coordinateur du second semestre de la spécialité informatique de la 3eme année de la formation initiale de Polytech'Paris-Sud.
- de septembre 2008 à août 2010, responsable des ter-stages en M1 BIBS.
- de septembre 2007 à septembre 2010, tuteur de Luidnel Maignan (moniteur en informatique de l'Université Paris-Sud).

Commissions de spécialistes

- 2010 : membre du comité de sélection d'AgroParisTech pour la campagne de recrutement des postes de maîtres de conférences.
- 2010, 2008 et 2007 : membre d'un des comités de sélection de la section 27 de l'Université Paris-Sud pour la campagne de recrutement des postes de maîtres de conférences.
- 2009 : membre du comité de sélection de l'ENSIIE[10] d'Evry pour la campagne de recrutement des postes de maîtres de conférences.

Autres responsabilités – vie du laboratoire

[10] École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise

- depuis septembre 2007, administrateur du site web du GDR CNRS 3003 - Bioinformatique Moléculaire.
- depuis septembre 2007, responsable des séminaires de l'équipe Bioinformatique.

Sommaire

Jury de Thèse
Comité de Programme de Conférences internationales

Animations scientifiques majeures

Jury de Thèse

- Examinateur de la thèse de Paola Salle soutenue le 13 juillet 2010 au LIRMM (Montpellier)
- Examinateur de la thèse de Guillaume Renvez soutenue le 25 juin 2010 au LAAS (Toulouse)
- Examinateur de la thèse de Laurent Brisson soutenue le 13 décembre 2006 au I3S (Nice - Sophia Antipolis)

Comité de Programme de Conférences internationales

- Membre des comités de programmes de DMIN, QIMIE, EGC, QDC et EvalECD.
- Co-organisateur des éditions 2009, 2010 et 2011 de l'atelier QDC (Qualité des Données et des Connaissances). Cet atelier est associé à la conférence EGC (Extraction et Gestion des Connaissances).
- Relecteur pour les revues TSI et RNTI.
- Relecteur pour DILS, DMIN, ECML/PKDD, CAP, CORIA, JOBIM, EGC, etc.

Sommaire

Participation
Création

Challenges internationaux de fouilles de données

Participation

- compétitions CAPRI[11] depuis 2005. Obtention d'un "High quality prediction" lors du round 16 (décembre 2008) dand la session *Scorer*. Obtention d'une solution "Acceptable" lors du round 26, target 53 (décembre 2011) dans la session *Predictor*.
- challenge "Novelty" de TREC[12] 2004. Résulat : $8^{ème}/55$
- challenge "Physiological Data Modeling Contest" d'ICML[13] 2004. Résulat : $7^{ème}/16$
- challenge "Atherosclérose" d'ECML/PKDD[14] 2002 (meilleur résultat)

Création

- Co-créateur et co-animateur des Challenges DEFT (éditions 2005 et 2006)
  Créateur avec Mathieu Roche du DÉfi Fouille de Textes DEFT'05. Ce défi a été organisé, lors de la première édition en 2005, sous la forme d'un atelier de la conférence TALN'05 (Traitement Automatique du Langage Naturel), puis en 2006, sous la forme d'un atelier de la conférence SDN (Semaine du Document Numérique). La gestion et l'animation de DEFT a été reprise par d'autres laboratoires, le LIMSI, le LIA, etc. La conférence TALN a accueilli les éditions 2010 et 2011 du challenge DEFT et devrait également accueillir l'édition 2012.

[11] **C**ritical **A**ssesment of **PR**ediction **I**nteractions
[12] **T**ext **RE**trieval **C**onference
[13] **I**nternational **C**onference on **M**achine **L**earning
[14] **E**uropean **C**onference on **M**achine **L**earning / European Conference on **P**rinciples and **P**ractice of **K**nowledge **D**iscovery in **D**atabases

Sommaire

Post-Doc
Thèse
Encadrement de Stage de Master Recherche

Encadrement de Post-Docs, de Thèses et de Masters Recherche

Post-Doc

- **Thomas Bourquard** (post-doc de janvier 2010 à juin 2010), [1,2,15,16,30,46]
- **Damien de Vienne** (post-doc de novembre 2008 à janvier 2010), article soumis [50]

Thèse

- **Julie Bernauer** (thèse soutenue le 7 avril 2006). Co-encadrement à partir de janvier 2004, à hauteur de 25% avec Anne Poupon. “Utilisation de la tessellation de Voronoï pour la modélisation des complexes protéine-protéine.”. Dans le cadre de cette thèse, j’ai collaboré avec Anne Poupon et Julie Bernauer sur les problématiques d’apprentissage supervisé à base d’algorithmes évolutionnaires pour apprendre des fonctions d’ordonnancement de complexes protéine-protéine [3,5,32].
- **Lucie Gentils** (thèse soutenue le 26 septembre 2008). Co-encadrement à partir du 1er janvier 2006, à hauteur de 50% avec Christine Froidevaux (encadrante principale). “Conception d’un module d’annotation semi-automatique de génomes à l’aide d’une hiérarchie fonctionnelle”. L’objectif de cette thèse était d’étudier et de mettre en œuvre un système d’annotation fonctionnelle semi-automatique de protéines. Ce travail a été réalisé dans le cadre du projet RAFALE[15] [4,24,31,47,48].
- **Thomas Bourquard** (thèse soutenue le 17 décembre 2009). Co-encadrement à 50% avec Anne Poupon (DR, équipe “Biology and Bioinformatics of Signalling Systems”, INRA de

[15] http ://www.lri.fr/RAFALE

Tours). “Exploitation des algorithmes génétiques pour la prédiction de structure de complexe protéine-protéine” L’objectif de cette thèse était de concevoir et mettre en œuvre un système performant de modélisation gros-grain des complexes protéine-protéine, ainsi qu’un système d’évaluation des complexes modélisés. Ce travail s’inscrivait dans la poursuite directe des travaux de thèse de J. Bernauer [16].

– **Adrien Guilhot-Gaudeffroy** Thèse débutée en octobre 2011. Co-encadrement à 45 % avec Christine Froidevaux (10%) et Julie Bernauer (45%). L’objectif de cette thèse est de concevoir et de mettre en œuvre une méthode permettant de réaliser l’ammarage 3D de complexes protéines-ARN.

Encadrement de Stage de Master Recherche

– **Abdelkader Hamadi** du 30/03/2010 au 17/09/2010, “Annotation fonctionnelle des protéines par apprentissage actif”
– **Adrien Guilhot-Gaudeffroy** du 30/03/2011 au 04/07/2011, poursuite en thèse, “Modélisation des complexes protéine-ARN”, [45]

Sommaire

Industrielles
Académiques

Collaborations

Industrielles

– Depuis septembre 2011, collaboration avec L'Oréal sur des thématiques de fouilles de données.

Académiques

– Depuis octobre 2011, collaboration avec Maguelonne Teisseirre (Cemagref Montpellier, IRSTEA-TETIS) [14,42]
– Depuis janvier 2004, collaboration avec Anne Poupon (Équipe "Biology and Bioinformatics of Signalling Systems", UMR PRC, INRA de Tours) [1,2,3,5,15,16,30,32,46]
– Depuis janvier 2006, collaboration avec Valentin Loux et Jean-François Gibrat (Unité "Mathématique, Informatique et Génome", INRA de Jouy-en-Josas) [4,24,31,47,48]

Sommaire

Revues internationales avec comité de lecture
Revues nationales avec comité de lecture
Chapitres de livres
Éditions d'ouvrages
Actes de conférences internationales avec comité de lecture
Workshops Internationaux
Actes de Challenges Internationaux
Actes de conférences nationales avec comité de lecture
Workshops Nationaux
Posters

PUBLICATIONS SCIENTIFIQUES

	International	National
Revues avec comité de lecture	6	4
Conférences avec comité de lecture	10	12
Chapitres de livres	2	–
Éditions d'ouvrages	–	1
Workshops, posters, challenges	8	6

Revues internationales avec comité de lecture

[1] Damien M. de Vienne and **Jérôme Azé**. Efficient Prediction of Co-Complexed Proteins Based on Coevolution. PLoS ONE 7(11) : e48728 doi :10.1371/journal.pone.0048728

[2] Sarel J. Fleishman *et al.* Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology, *Journal of molecular biology (Impact Factor : 4.008)*, doi :10.1016/j.jmb.2011.09.031, 2011

[3] Thomas Bourquard, Julie Bernauer, **Jérôme Azé** and Anne Poupon A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions, *PLoS ONE (Impact Factor : 4.351)*, 6(4) :e18541, 2011

[4] Julie Bernauer, **Jérôme Azé**, Joël Janin, and Anne Poupon. A new protein-protein docking scoring function based on interface residues properties, *Bioinformatics (Impact Factor : 5.04)*, 5(23) :555-62, 2007.

[5] **Jérôme Azé**, Lucie Gentils, Claire Toffano-Nioche, Valentin Loux, Jean-François Gibrat, Philippe Bessières, Céline Rouveirol, Anne Poupon, and Christine Froidevaux. Towards a semi-automatic

functionnal annotation tool based on decision tree techniques, *BMC Proceedings*, 2(Suppl 4) :S3 2007.

[6] Julie Bernauer, Anne Poupon, **Jérôme Azé** and Joël Janin. A docking analysis of the statistical physics of protein-protein recognition, *Journal Physical Biology*, 2(2) :S17-23, 2005.

[7] **Jérôme Azé**, Mathieu Roche, Yves Kodratoff and Michèle Sebag. Learning to order terms : supervised interestingness measures in terminology extraction *In IJCI (International Journal on Computational Intelligence)*, 1(2) :104-107, http ://www.ijci.org/product/1304-2386-1.pdf. Revised version of proceedings of ICCI (International Conference on Computational Intelligence), p478-481, december 2004, Istanbul, Turkey.

Revues nationales avec comité de lecture

[8] **Jérôme Azé** et Yves Kodratoff. Extraction de "pépites" de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit. *Numéro spécial de la revue RNTI "Mesures de qualité pour la fouille de données", sous la direction de : H. Briand, M. Sebag, R. Gras, F. Guillet*, E-1 :247-270, avril 2004.

[9] Israël César Lerman et **Jérôme Azé**. Indice Probabiliste Discriminant de Vraisemblance du Lien pour des Données Volumineuses. *Numéro spécial de la revue RNTI "Mesures de qualité pour la fouille de données", sous la direction de : H. Briand, M. Sebag, R. Gras, F. Guillet*, E-1 :69-94, avril 2004.

[10] **Jérôme Azé**, Sylvie Guillaume et Philippe Castagliola. Évaluation de la Résistance au Bruit de quelques Mesures Quantitatives. *Dans les actes des XXXVièmes Journées de Statistique (résumé) Volume 1, pp 133-136 et dans le numéro spécial de la revue RNTI "Entreposage et Fouille de données"*, 1 :159–170, 2003.

[11] Yves Kodratoff, **Jérôme Azé**, Mathieu Roche et Oriane Matte-Tailliez. Des textes aux associations entre les concepts qu'ils contiennent. *Dans les actes des XXXVièmes Journées de Statistique (résumé) Volume 2, p599-602 et dans le numéro spécial de la revue RNTI "Entreposage et Fouille de données"*, 1 :171–182, 2003.

Chapitres de livres

[12] Yves Kodratoff et **Jérôme Azé** et Lise Fontaine. CorTag : a contextual tagging of words within their sentences. in Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration, Violaine Prince and Mathieu Roche (eds), London : IGI Publishing, 177-189, 2009

[13] Israël César Lerman et **Jérôme Azé**. A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link. *Springer*, 207-236, 2007.

Éditions d'ouvrages

[14] Mesurer et évaluer la qualité des données et des connaissances. *Rédacteurs invités :* **Jérôme Azé**, *Nicolas Béchet, Laure Berti-Équille, Sylvie Guillaume, Mathieu Roche et Fatiha Saïs*. Eds Hermann,

RNTI-E-22, ISBN : 9782705682866, 2011

Actes de conférences internationales avec comité de lecture

[15] Hugo Alatrista Salas, **Jérôme Azé**, Sandra Bringay, Flavie Cernesson, Frédéric Flouvat, Nazha Selmaoui-Folcher and Maguelonne Teisseire. Finding Relevant Sequences With The Least Temporal Contradiction Measure : Application to Hydrological Data *the 15th AGILE International Conference*. to be publish, 2012

[16] **Jérôme Azé**, Thomas Bourquard, Sylvie Hamel, Anne Poupon and Dave Ritchie. Using Kendall-Tau Meta-Bagging to Improve Protein-Protein Docking Predictions *The 6th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB'2011)*, pp. 284-295, 2011.

[17] Thomas Bourquard, Julie Bernauer, **Jérôme Azé**, and Anne Poupon. Comparing Voronoi and Laguerre tessellations in the protein-protein docking context, *Sixth International Symposium on Voronoi Diagrams (ISVD'2009)*, pp. 225-232, 2009.

[18] **Jérôme Azé**, Philippe Lenca, Stéphane Lallich, and Benoît Vaillant. A study of the robustness of association rules In *The 2007 International Conference on Data Mining (DMIN'07)*, pp 132-137, 2007

[19] **Jérôme Azé**. A clustering approach for analysing association rules. In *ASMDA'07, special session "Quality Measures in Data Mining", International Symposium on "Applied Stochastic Models and Data Analysis"*, 2007.

[20] **Jérôme Azé**, Mathieu Roche, Yves Kodratoff, and Michèle Sebag. Preference Learning in Terminology Extraction : A ROC-based approach. In *ASMDA'05, International Symposium on "Applied Stochastic Models and Data Analysis"*, 17-20 mai 2005, Brest, France.

[21] Mathieu Roche, **Jérôme Azé**, Oriane Matte-Tailliez, and Yves Kodratoff. Mining texts by association rules discovery in a technical corpus. In *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, pp 89-98, 17-20 may 2004, Zakopane, Poland.

[22] Michèle Sebag, **Jérôme Azé**, and Noël Lucas. Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning. In *Proceedings of the Third IEEE International Conference on Data Mining, ICDM 2003*, pp 637-640, 19-22 November 2003, Melbourne, Florida, USA.

[23] Michèle Sebag, Noël Lucas, and **Jérôme Azé**. ROC-based evolutionary learning : Application to medical data mining. In *Proceedings of the 6th International Conference on Artificial Evolution, EA 2003*, pp 384-396, 27-30 October 2003, Marseille, France.

[24] **Jérôme Azé** and Yves Kodratoff. A study of the Effect of Noisy Data in Rule Extraction Systems. In *Proceedings of the Sixteenth European Meeting on Cybernetics and Systems Research, EMCSR'02*, volume 2, pp 781–786, 2-5 april 2002, Vienna, Austria.

Workshops Internationaux

[25] **Jérôme Azé**, Lucie Gentils, Claire Toffano-Nioche, Valentin Loux, Jean-François Gibrat, Philippe Bessières, Céline Rouveirol, Anne Poupon, and Christine Froidevaux. Towards a semi-automatic functionnal annotation tool based on decision tree techniques. In *Proceedings of the International Workshop on Machine Learning in Systems Biology (MSLB'07)*, 2007.

[26] Mathieu Roche, **Jérôme Azé**, Yves Kodratoff, and Michèle Sebag. Learning Interestingness Measures in Terminology Extraction. A ROC-based approach. In *Proceedings of "ROC Analysis in AI" Workshop ECAI 2004 (European Conference on Artificial Intelligence 2004), Valencia, Spain*, pp 81–88, 21 august 2004, Valencia, Spain.

Actes de Challenges Internationaux

[27] Ahmed Amrani, **Jérôme Azé**, Thomas Heitz, Yves Kodratoff, and Mathieu Roche. From the texts to the concepts they contain : a chain of linguistic treatments. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, 16-19 november 2004, Gaitherburg, Maryland, USA.

[28] **Jérôme Azé**, Noël Lucas, and Michèle Sebag. A Genetic ROC-based Classifier In *Proceedings of "Physiological Data Modeling Contest", workshop of the The Twenty-First International Conference on Machine Learning (ICML 2004)*, 8 july 2004 Banff, Alberta, Canada.

[29] **Jérôme Azé**, Noël Lucas, and Michèle Sebag. A new medical test for atherosclerosis detection : Geno. In *Proceedings of Discovery Challenge, workshop of The 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003)*, 22-26 september 2003, Cavtat-Dubrovnik, Croatia.

[30] Noël Lucas, **Jérôme Azé**, and Michèle Sebag. Atherosclerosis risk identification and visual analysis. In *Proceedings of Discovery Challenge, workshop of The 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2002)*, 19-23 august 2002, Helsinki, Finland.

Actes de conférences nationales avec comité de lecture

[31] Thomas Bourquard, **Jérôme Azé**, Anne Poupon and David W. Ritchie. Amarrage protéine-protéine par couplage de la complémentarité de forme et des empreintes Voronoï. 12èmes Journées Ouvertes Biologie Informatique Mathématiques (JOBIM 2011), pp 9-16, 2011.

[32] Lucie Gentils, **Jérôme Azé**, Claire Toffano-Nioche, Valentin Loux, Anne Poupon, Jean-François Gibrat, and Christine Froidevaux. Mesures hiérarchiques pondérées pour l'évaluation d'un système semi-automatique d'annotation de génomes utilisant des arbres de décision. Revue *RNTI*, Numéro spécial conférence Extraction et Gestion des Connaissances (EGC'08), 1 :133-138, 2008.

[33] Julie Bernauer, **Jérôme Azé**, Joël Janin and Anne Poupon. Une nouvelle fonction de score pour l'amarrage protéine-protéine fondée sur les diagrammes de Voronoï, In proceedings of Journés Ouvertes de Biologie Informatique Mathématiques, Lyon, France. 459-469, 2005

[34] Ahmed Amrani, **Jérôme Azé** et Yves Kodratoff. Logiciel d'aide à l'étiquetage morpho-syntaxique de textes de spécialité. Revue *RNTI*, Numéro spécial conférence EGC'05, RNTI-E-3(Vol II) :673-678, 19-21 Janvier 2005, Paris, France.

[35] Israël César Lerman et **Jérôme Azé**. Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. Revue *RSTI série RIA-ECA*, Numéro spécial conférence EGC'03, 17(1-2-3) :247–262, 22-24 janvier 2003, Lyon, France.

[36] **Jérôme Azé**. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. Revue *RSTI série RIA-ECA*, Numéro spécial conférence EGC'03, 17(1-2-3) :171–182, 22-24 janvier 2003, Lyon, France.

[37] **Jérôme Azé** et Mathieu Roche. Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. Revue *RSTI série RIA-ECA*, Numéro spécial conférence EGC'03, 17(1-2-3) :283–294, 22-24 janvier 2003, Lyon, France.

[38] **Jérôme Azé**, Noël Lucas et Michèle Sebag. Fouille de données visuelle et analyse de facteurs de risque médical. Revue *RSTI série RIA-ECA*, Numéro spécial conférence EGC'03, 17(1-2-3) :183–188, 22-24 janvier 2003, Lyon, France.

[39] Mathieu Roche, Oriane Matte-Tailliez, **Jérôme Azé** et Yves Kodratoff. Extraction de la Terminologie du Domaine : Étude de Mesures sur un Corpus Spécialisé Issu du Web. In *Proceedings of JFT'03 (Journées Francophones de la Toile 2003)*, pp 279–288, 30 juin et 1-2 juillet 2003, Tours, France.

[40] Michèle Sebag, **Jérôme Azé** et Noël Lucas. Apprendre et optimiser la courbe ROC. In *Proceedings of Conférence d'Apprentissage (CAp 2003), Plate-forme AFIA 2003*, pp 315–330, 1-2 juillet 2003, laval, France.

[41] **Jérôme Azé** et Yves Kodratoff. Évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. *Revue RSTI-ECA (Numéro spécial EGC'2002)*, 1(4) :143–154, 21-23 janvier 2002, Montpellier, France.

[42] **Jérôme Azé**, Hélène Richy, and Guy Lorette. Interface-stylo de correction en ligne de documents électroniques : Application aux pages web. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED'2000)*, Hubert Emptoz, Nicole Vincent (eds.), pp 261-269, 3-5 juillet 2000, Lyon, France.

Workshops Nationaux

[43] Hugo Alatrista Salas, **Jérôme Azé**, Flavie Cernesson, Sandra Bringay and Maguelonne Teisseire. Fouille de motifs appliquée à des données hydrologiques *Atelier EvalECD 2011, en conjonction avec EGC 2011*, 2011

[44] Yves Kodratoff, Oriane Matte-Tailliez, Mathieu Roche, **Jérôme Azé** et Ahmed Amrani. Une chaîne complète pour le traitement de quantités de textes : des textes bruts à l'extraction de l'information. In *Actes de l'atelier "Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel" à la plateforme AFIA 2003, 4 juillet 2003, Laval*, France.

[45] Oriane Matte-Tailliez, Mathieu Roche, **Jérôme Azé** et Yves Kodratoff. Extraction automatique de la terminologie en biologie moléculaire : une étape fondamentale pour l'extraction d'information. In *Actes de l'atelier "Fouille de textes en génomique" à EGC'2003*, 22 Janvier 2003, Lyon, France.

Posters

[46] Adrien Guilhot-Gaudeffroy, **Jérôme Azé**, Julie Bernauer. Prediction of protein-nucleic acid interactions : building a scoring function with the Voronoi diagram. In *Stanford-Sweden multiresolution Molecular simulation workshop*, poster, June 2011

[47] Thomas Bourquard, **Jérôme Azé**, Anisah W. Ghoorah, Anne Poupon, Dave W. Ritchie. Using Voronoï fingerprints to rescore Hex protein-protein docking models. In *Colloquium Bioinformatique du LIX*, poster, du 8 au 10 novembre 2010.

[48] Lucie Gentils, **Jérôme Azé**, Philippe Bessières, Jean-François Gibrat, Valentin Loux, Céline

Rouveirol, and Christine Froidevaux. Learning rules for microbial genome annotation, In *7èmes Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)*, poster, 73-74, 2006.

[49] Lucie Gentils, **Jérôme Azé**, Philippe Bessières, Jean-François Gibrat, Valentin Loux, Céline Rouveirol, and Christine Froidevaux. Learning rules for microbial genome annotation, In *Automated Function Prediction* poster, 82-83, San Diego, CA, USA, 2006.

[50] Yves Kodratoff, Oriane Matte-Tailliez, Mathieu Roche, **Jérôme Azé** et Ahmed Amrani. Extraction de règles d'association de concepts à partir de différents corpus spécialisés, plateforme AFIA 2003, 4 juillet 2003, Laval, France.