NOTICE DE TITRES ET TRAVAUX

Yann Guermeur

8 janvier 2015

1 Curriculum vitae

1.1 Etat civil et coordonnées

Nom : Guermeur
Prénoms : Yann, Charles, Louis, Marie
Date et lieu de naissance : 19 mai 1967, à Neuilly-sur-Seine, Hauts-de-Seine (92)
Nationalité : Française
Situation de famille : célibataire

Coordonnées professionnelles

LORIA, équipe ABC, Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy cedex, France

Téléphone : +33 3 83 59 30 18
Télécopie : +33 3 83 27 83 19

Adresse électronique : Yann.Guermeur@loria.fr
Page web : http://www.loria.fr/~guermeur/

Fonctions et établissement actuels

Directeur de recherche (DR2) au CNRS, affecté au LORIA - UMR 7503

Responsable scientifique de l'équipe "Apprentissage et Biologie Computationnelle" (ABC) du LORIA : http://abc.loria.fr/

1.2 Titres universitaires

- 2007
Habilitation à Diriger des Recherches (HDR) en Informatique de l'Université Henri Poincaré (UHP) - Nancy 1
Parrain scientifique : Jean-Paul Haton
Titre : SVM multiclasses, théorie et applications
Date de soutenance : 28 novembre 2007
Lieu de soutenance : LORIA
Composition du jury : Monsieur Tombre, Président, Monsieur Denis, Rapporteur, Monsieur Boucheron, Rapporteur, Monsieur Gascuel, Rapporteur, Monsieur Haton, Parrain scientifique, Madame Sebag, Examinateur, Monsieur Gallinari, Examinateur

- 1997
Doctorat de l'Université Paris 6, spécialité Informatique, mention très honorable avec félicitations
Directeur de thèse : Patrick Gallinari
Titre : Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines

Date de soutenance : 10 décembre 1997
Lieu de soutenance : Université Paris 6
Composition du jury : Monsieur Lebbe, Président, Madame Paugam-Moisy, Rapporteur, Monsieur Deléage, Rapporteur, Monsieur Gallinari, Directeur de thèse, Madame d'Alché-Buc, Examinateur, Monsieur Nadal, Examinateur

- 1993
DEA "Intelligence Artificielle, Reconnaissance des Formes et Applications" (IARFA), de l'Institut Blaise Pascal (IBP), Université Paris 6, mention bien

- 1991
Diplôme d'ingénieur de l'Institut d'Informatique d'Entreprise (IIE), école du concours commun Centrale-Supélec (l'IIE est devenu en 2006 l'Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE) et appartient à présent au concours commun Mines-Ponts)

- 1991
"Certificate of Proficiency in English" de l'Université de Cambridge

1.3 Stages, expérience professionnelle

- D'octobre 2003 à ce jour

Chercheur au CNRS (initialement CR1, DR2 depuis octobre 2011), affecté au LORIA. Membre de l'équipe MODBIO jusqu'au 30 juin 2006, responsable scientifique de l'équipe ABC depuis cette date.

- De février 1999 à septembre 2003

Maître de conférences à l'UHP, affecté au site de Saint-Dié de l'IUT de Nancy Brabois, devenu l'IUT de Saint-Dié le premier janvier 2000. La ville de Saint-Dié se nomme à présent Saint-Dié-des-Vosges. Membre de l'équipe CORTEX du LORIA, dirigée par Frédéric Alexandre, jusqu'au 31 décembre 2000. Membre de l'équipe MODBIO, dirigée par Alexander Bockmayr, après cette date.

- De septembre 1998 à janvier 1999

Chercheur post-doctoral dans l'équipe connexionniste dirigée par Patrick Gallinari, au sein du thème Apprentissage et Acquisition des Connaissances - APprentissage Automatique (APA) du Laboratoire d'Informatique de Paris 6 (LIP6), à l'Université Paris 6. Durant ce stage, j'ai participé à la conception d'un système de modélisation statistique des utilisateurs de l'Internet.

- De septembre 1997 à août 1998

ATER à l'Ecole Normale Supérieure de Lyon, membre de l'équipe connexionniste du Laboratoire de l'Informatique du Parallélisme (LIP) dirigée par Hélène Paugam-Moisy.

- De septembre 1996 à août 1997

ATER à l'UFR d'Informatique de l'Université Paris 7.

- D'octobre 1994 à août 1996

Moniteur à l'UFR d'Informatique de l'Université Paris 7.

- De mars à mai 1994

Vacataire à l'Institut de Statistique de l'Université Pierre et Marie Curie (ISUP).

- De novembre 1993 à août 1996

Doctorant, allocataire MRE, sous la direction de Patrick Gallinari, dans l'équipe connexionniste du LAFORIA, puis du LIP6, à l'Université Paris 6.

- D'avril à septembre 1993

Stage de DEA effectué dans l'équipe connexionniste du LAFORIA, à l'Université Paris 6, sous la direction de Patrick Gallinari. Sujet : "Identification du locuteur, prise en compte de la dynamique par des systèmes récurrents".

- De septembre 1991 à août 1992

Société CEGI, à Paris. Responsable d'une application de gestion destinée aux associations "Profession Sport" dépendant du ministère de la Jeunesse et des Sports.

- De janvier à juin 1991

Institut Polytechnique de Leeds (Grande-Bretagne). Stage de recherche effectué sous la direction de Nairn Kennedy, dans le cadre du programme ERASMUS. Rédaction du mémoire de fin d'études de l'IIE intitulé "Formation de concepts dans les réseaux neuronaux".

- De juin à octobre 1990

Société EDGETEK, Département Ingénierie, aux Ulis. Stage portant sur la conception et le développement d'un logiciel de test de composants électroniques.

- De juillet à septembre 1989

AEROSPATIALE, Division Engins Tactiques, à Châtillon. Stage consacré à l'analyse et à la réalisation d'un logiciel de gestion des affaires traitées par la division.

2 Enseignement et diffusion de la culture scientifique

2.1 Tableau synthétique des enseignements effectués

Le tableau ci-dessous résume les enseignements que j'ai effectués à ce jour. L'ensemble représente 333 heures de cours magistraux (Cours), 878 heures de travaux dirigés (TD) et 137 heures de travaux pratiques (TP).

Etablissement	Enseignement	Activité	Etudiants	Vol. horaire
UHP	Bioinformatique	Cours	M2P Génomique et Info.	24
INPL	Bioinformatique	Cours	troisième année	6
UHP	Bioinformatique	Cours	DESS RGTI	20
UHP	Système (UNIX)	Cours	DESS RGTI	6
UHP	Bioinformatique	Cours	Maîtrise biol. (MBCP)	29
UHP	Bioinformatique	Cours	DEA Info.	6
UHP	Système (UNIX)	Cours	DUT SRC 2	32
UHP	Système (UNIX)	Cours	DUT SRC 1	50
UHP	Stat. et probabilités	Cours	DUT SRC 2	92
UHP	Programmation en C	Cours	DUT SRC 2	2
UHP	Programmation en C	Cours	DUT SRC 1	4
UHP	Programmation en JAVA	Cours	DUT SRC 1	34
ENS	Connexionnisme	Cours	Magistère Info. et Math. 2	16
ENS	Connexionnisme	Cours	Maîtrise Sciences Co.	12
UHP	Bioinformatique	TD	M2P Génomique et Info.	16
UHP	Bioinformatique	TD	Maîtrise biol. (MBCP)	4
UHP	Bases de données (DBASE)	TD	DEUG SV2	60
UHP	Programmation en CAML	TD	DEUG MIAS1	28
UHP	Système (UNIX)	TD	DUT SRC 2	60
UHP	Stat. et probabilités	TD	DUT SRC 2	198
UHP	Programmation en C	TD	DUT SRC 2	44
UHP	Programmation en C	TD	AETP	6
UHP	Programmation en C	TD	DUT SRC 1	24
UHP	Système (UNIX)	TD	DUT SRC 1	80
UHP	Programmation en JAVA	TD	DUT SRC 1	48
ENS	Connexionnisme	TD	Licence Sciences Co.	16
ENS	Fonctionnement (Système)	TD	Magistère Info. et Math. 1	24
ENS	Programmation en C	TD	Maîtrise Sciences Co.	14
P7	Réseaux	TD	Maîtrise d'Informatique	24
P6	Informatique et Programmation	TD	ISUP, première année	32
P7	Programmation en Pascal	TD	DEUG Sciences	200
UHP	Bioinformatique	TP	M2P Génomique et Info.	10
UHP	Programmation en C	TP	DUT SRC 1	40
UHP	Système (UNIX)	TP	DUT SRC 2	8
UHP	Programmation en HTML	TP	DUT SRC 1	32
UHP	Bioinformatique	TP	DESS RGTI	8
UHP	Système (UNIX)	TP	DESS RGTI	4
UHP	Bioinformatique	TP	Maîtrise biol. (MBCP)	35

TABLE 1 – ENS : Ecole Normale Supérieure de Lyon, INPL : Institut National Polytechnique de Lorraine, P6 : Université Paris 6, P7 : Université Paris 7, UHP : Université Henri Poincaré-Nancy 1.

Les cours donnés à l'INPL l'ont été dans le cadre de la formation de bioinformatique commune à l'Ecole Nationale Supérieure d'Agronomie et des Industries Alimentaires (ENSAIA), l'Ecole Nationale Supérieure des Industries Chimiques (ENSIC) et l'Ecole Nationale Supérieure des Mines de Nancy (ENSMN).

2.2 Détail des enseignements effectués

En tant qu'enseignant-chercheur, j'ai enseigné les principales matières de base de l'informatique : algorithmique, programmation, système, réseaux, bases de données... Ces enseignements ont été dispensés à l'ensemble des publics universitaires (élèves de grandes Ecoles, étudiants de facultés et d'IUT) en premier, deuxième et troisième cycle. Il s'agissait essentiellement d'étudiants en informatique et en biologie. Dès cette époque, j'ai également donné des cours en lien plus direct avec mon domaine de recherche : statistique et probabilités, apprentissage automatique et bioinformatique. Devenu chercheur, j'ai concentré mon activité d'enseignement sur la présentation des fondements et méthodes de l'apprentissage statistique, avec comme application privilégiée le traitement des données biologiques. Ainsi, de l'année universitaire 05-06 à l'année universitaire 08-09, j'ai été responsable de l'UE 3.105 "Apprentissage statistique et fouille de données" de la spécialité "Génomique et Informatique" en deuxième année du parcours professionnel (M2P) du Master "Sciences de la Vie et de la Santé" (SVS) de l'UHP. Le support du cours que j'ai donné dans ce cadre est disponible à l'adresse suivante : `http://www.loria.fr/~guermeur/Cours_08.pdf`.

2.3 Enseignements dans des écoles d'été

En juillet 2008, j'ai donné deux cours dans le cadre de la "Summer School on Neural Networks in Classification, Regression and Data Mining" (NN 2008, `http://www.isep.ipp.pt/nn/`). Les supports de ces cours de deux heures, intitulés respectivement "Multi-class Support Vector Machines" et "Protein Secondary Structure Prediction with Multi-class Support Vector Machines", sont disponibles en ligne aux adresses suivantes : `http://www.loria.fr/~guermeur/NN2008_M_SVM_YG.pdf` et `http://www.loria.fr/~guermeur/NN2008_Struct_YG.pdf`.

2.4 Communications invitées dans des conférences et séminaires à l'étranger

Le 26 mars 1999, j'ai donné au département d'informatique de la "Royal Holloway, University of London" (RHUL), un séminaire intitulé "A hierarchical method for protein secondary structure prediction" (`http://www.cs.rhul.ac.uk/Outreach/News-and-Events/1999/eventsarchives.htm`).

Le 20 septembre 2000, j'ai donné au département d'informatique et des sciences de l'information de l'Université de Gênes un séminaire intitulé "A new SVM for multi-category discriminant analysis" (`http://www.disi.unige.it/index.php?eventsandseminars/old-disi-seminars`).

Le 14 avril 2003, j'ai présenté une communication dans le cadre du workshop "Kernel Methods in Computational Biology" de la "Seventh Annual International Conference on Research in Computational Molecular Biology" (RECOMB'03). Son titre était "A hybrid kernel machine for protein secondary structure prediction".

Le 4 juillet 2003, j'ai donné une conférence dans le cadre de l'atelier "Apprentissage Machine et Bioinformatique" de la plateforme AFIA 2003. Cette conférence était intitulée "Analyse de séquence et prédiction de structure de protéines" (`http://afia.lri.fr/plateforme-2003/Programme/Vendredi-4.html`).

Le 24 octobre 2005, j'ai présenté une communication dans le cadre de la session thématique 2 bio-informatique du congrès ASTI'2005. Celle-ci avait pour titre "Apprentissage automatique, applications en prédiction de la stucture secondaire et tertiaire des protéines" (`http://www.isima.fr/asti2005/ann/appr_gen.pdf`).

Le 9 avril 2008, à l'occasion des troisièmes journées thématiques "Apprentissage Artificiel & Fouille de Données" (AAFD'08), j'ai effectué un exposé intitulé "Risques garantis et sélection de

modèle pour les SVM multi-classes, application à la prédiction de la structure secondaire des protéines" (`http://www-lipn.univ-paris13.fr/A3/AAFD08/programme.html`).

Le 29 août 2008, j'ai présenté dans la session "Machines à vecteurs de support et autres méthodes à noyaux" des journées "Modélisation Aléatoire et Statistique" (MAS) de la "Société de Mathématiques Appliquées et Industrielles" (SMAI) une communication intitulée "Risques garantis pour les M-SVM" (`http://mas2008.univ-rennes1.fr/download/slide.php`).

Le 17 juin 2010, j'ai donné une conférence plénière intitulée "Linear ensemble methods for multiclass support vector machines" dans le cadre du workshop "Optimization and Learning : Theory, Algorithms and Applications" (WS'10) (`http://www.lita.univ-metz.fr/%7Ews10/index.php`).

3 Encadrement d'activités de recherche

3.1 Stages de DEA - Master 2 recherche

Durant l'année scolaire 97-98, Hélène Paugam-Moisy et moi avons co-encadré Olivier Teytaud, élève normalien effectuant son stage du DEA Informatique de Lyon (DIL). Olivier a rédigé un mémoire intitulé "Représentations internes dans les réseaux de neurones artificiels". Ce travail a produit des résultats originaux dans deux domaines : le calcul des dichotomies polyédriques par les PMC à unités à seuil et l'étude des capacités de généralisation des SVM. Olivier est actuellement chargé de recherche à l'INRIA.

Durant l'année scolaire 01-02, j'ai encadré Régis Vert, élève de l'Ecole Nationale Supérieure des Mines de Nancy (ENSMN) effectuant son stage du DEA Informatique de Lorraine. Le sujet du stage était la conception et la mise en œuvre de M-SVM dédiées au traitement de séquences biologiques. Le principal résultat de cette étude a été une extension du principe d'alignement noyau-cible au cas multi-classe. Ceci a permis à Régis de proposer un nouveau noyau pour la prédiction de la structure secondaire des protéines (voir en particulier la référence [CL02] de ma liste de publications). Après avoir soutenu une thèse en théorie statistique de l'apprentissage à l'Université Paris 11, en juin 2006, Régis a effectué un stage post-doctoral au "Max-Planck-Institut für biologische Kybernetik", à Tübingen. Il est à présent chercheur dans le privé.

Durant l'année scolaire 02-03, j'ai encadré le stage du DEA Informatique de Lorraine d'Emmanuel Didiot. Celui-ci a poursuivi les recherches de Régis Vert sur la mise au point d'un noyau dédié au traitement des séquences protéiques. La validation de ces travaux a de nouveau été effectuée en prédiction de la structure secondaire des protéines. Après avoir soutenu une thèse en parole dans l'équipe Parole du LORIA, en novembre 2007, Emmanuel a été membre de l'équipe ABC durant l'année scolaire 07-08, en tant qu'ATER. Il occupe à présent un emploi d'ingénieur au LORIA. De mai à juillet 2003, j'ai également encadré Sumit Kumar Jha, étudiant en avant-dernière année de l'IIT de Kharagpur, en Inde. Nous avons travaillé sur l'application d'une M-SVM à la prédiction des ponts disulfures. Sumit a effectué sa thèse à l'Université Carnegie Mellon (CMU), où il a conservé comme domaine d'application de ses recherches la prédiction du repliement des protéines. Diplômé en juillet 2010, il est à actuellement "Assistant Professor" à l'Université de Floride Centrale, à Orlando.

Durant l'année scolaire 05-06, Nadir T. Mrabet et moi avons co-encadré Levoly Fani, élève de l'Ecole Nationale Supérieure d'Electricité et de Mécanique (ENSEM), à l'occasion de son stage du Master Informatique, dans la spécialité "Services Distribués et Réseaux de Communication" (SDRC), proposée par l'UHP, l'Université Nancy 2 et l'INPL. Levoly a rédigé un mémoire intitulé "Spiralix, un programme PERL pour définir les positions frontières d'hélices α : optimisation et représentation vectorielle des hélices α". Il travaille à présent dans le privé.

Durant l'année scolaire 06-07, j'ai encadré deux étudiants en stage de deuxième année de Master recherche. Aurélie Colas, élève de l'ENSMN, suivait les enseignements du Master Chimie et Physico-chimie Moléculaires (CPM) dans la spécialité "Chimie Informatique et Théorique" (CIT), de l'UHP et de l'INPL. L'intitulé de son travail de stage était : "Mise en œuvre d'une solution efficace au problème de programmation quadratique de grande taille correspondant à l'apprentissage des SVM multiclasses". Julien Vannesson, étudiant du Master Informatique, dans la spécialité "Perception, Raisonnement, Interaction Multimodale" (PRIM), proposée par l'UHP, l'Université Nancy 2 et l'INPL, a travaillé à l'exploitation par notre méthode de prédiction de la structure secondaire des protéines des alignements multiples. De septembre à novembre 2007, il a poursuivi ses recherches comme ingénieur CNRS. Aurélie et Julien travaillent à présent dans le privé.

Durant l'année scolaire 13-14, Fabien Lauer et moi avons co-encadré Khadija Musayeva, élève en deuxième année du Master mention informatique de l'Université de Lorraine (UL). Ses travaux portaint sur l'"Apprentissage statistique pour la segmentation de séquences biologiques". Khadija a pu les poursuivre d'octobre à décembre 2014, comme ingénieur de recherche dans l'équipe ABC.

3.2 Thèses

De septembre 2003 à septembre 2006, Alexander Bockmayr et moi avons co-encadré le travail de thèse de Yannick Darcy intitulé "Conception, mise en œuvre et évaluation de machines à noyau dédiées au traitement de séquences biologiques". Le financement de cette thèse de l'UHP par une bourse ministérielle a été obtenu dans le cadre du projet GENOTO3D financé par l'ACI "Masses de Données" (voir la section 4.1). Des raisons personnelles ont conduit Yannick à décider d'interrompre ses recherches pour partir à l'étranger. Celles-ci ont fait l'objet de publications et ont été poursuivies par d'autres membres de l'équipe ABC.

De mai 2008 jusqu'à la soutenance, le 27 novembre 2014, j'ai co-encadré avec Belhadri Messabih la thèse d'Hafida Chouarfia, maître assistante chargée de cours au Département d'Informatique de l'Université des Sciences et de la Technologie d'Oran (USTO MB). Cette thèse de l'USTO portait sur la "Prédiction de la structure protéique". Hafida a effectué un séjour de dix-huit mois au LORIA d'octobre 2009 à mars 2011, dans le cadre du Programme Franco-Algérien de Formation Supérieure (PROFAS).

De septembre 2009 au 19 juin 2013, date de la soutenance, j'ai dirigé avec Samy Tindel le travail de thèse de Rémi Bonidal intitulé "Sélection de modèle par chemin de régularisation pour les machines à vecteurs support à coût quadratique". Les trois premières années de cette thèse de l'Université de Lorraine ont été co-financées par la Fédération Charles Hermite (`http://www.fr-hermite.univ-lorraine.fr/`) et la Région Lorraine. Durant l'année universitaire 2012-2013, Rémi a occupé un poste d'ATER à l'UFR Mathématiques et Informatique de l'Université de Lorraine. Il travaille à présent dans l'industrie, à Lyon.

Depuis décembre 2010, je co-encadre avec Abdelkader Benyettou la thèse de l'USTO de Mounia Hendel, maître assistante chargée de cours à l'Ecole Préparatoire en Sciences et Techniques d'Oran (EPSTO). Ce travail de recherche est intitulé "Etude des M-SVM appliquées aux signaux physiologiques".

De mars 2011 au 21 novembre 2013, date de la soutenance, j'ai dirigé avec Matthieu Geist la thèse d'Edouard Klein intitulée "Contributions à l'apprentissage par renforcement inverse". Cette thèse de l'Université de Lorraine a été entièrement financée par Supélec. A l'issue d'une année sabbatique, Edouard a rejoint une unité de recherche de la gendarmerie, avec le grade de chef d'escadron.

3.3 Travaux d'ingénieurs de recherche

Julien Vannesson et Khadija Musayeva ont tout deux bénéficié d'un support d'ingénieur de recherche de trois mois afin d'achever les travaux qu'ils avaient débutés dans le cadre de leur stage de Master 2 (voir la section 3.1).

Depuis septembre 2013, j'encadre les travaux d'Emmanuel Didiot, ingénieur de recherche au CNRS. Ce financement de 30 mois accordé par l'INS2I du CNRS a été attribué au projet "Apprentissage statistique pour la segmentation de séquences biologiques" (A3SB) dont je suis le porteur.

3.4 Recherches post-doctorales

D'octobre 2004 à août 2005, j'ai encadré le stage post-doctoral du STIC-CNRS de Frédéric Sur, traitant du sujet suivant : "Modèles statistiques hybrides pour la recherche de motifs dans les séquences biologiques". Frédéric Sur occupe à présent un poste de maître de conférences à l'Université de Lorraine (il est affecté à l'ENSMN).

D'octobre 2005 à mars 2007, j'ai supervisé le stage post-doctoral qu'a effectué Emmanuel Monfrini dans le cadre du projet "Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques" du programme Décrypthon (voir la section 5.1). Ce stage, financé par l'Association Française contre les Myopathies (AFM), portait sur la conception et le développement d'un noyau dédié à l'identification des différentes catégories intervenant dans les phénomènes d'épissage alternatif. Pendant les six premiers mois du projet, les bases de données utilisées par Emmanuel ont été assemblées par un ingénieur d'études, Delphine Autard. Celle-ci occupe à présent un poste d'ingénieur d'études permanent à l'INSERM. Depuis septembre 2008, Emmanuel Monfrini est maître de conférences à l'Institut TELECOM.

4 Administration de la recherche et responsabilités collectives

4.1 Actions nationales et internationales

De février 1999 à février 2003, j'ai participé au groupe de travail ESPRIT "Neural and Computational Learning Theory" (NeuroCOLT2) : `http://www.neurocolt.com/`. Ce groupe de travail s'est poursuivi jusqu'en février 2008, à travers le réseau d'excellence européen "Pattern Analysis, Statistical Modelling and Computational Learning" (PASCAL, `http://www.pascal-network.org/`), puis jusqu'en février 2013, à travers le réseau PASCAL 2 (`http://pascallin2.ecs.soton.ac.uk/`). J'ai été membre du comité d'organisation du challenge théorique PASCAL "Type I and type II errors for multiple simultaneous hypothesis testing" (`http://www.lri.fr/~teytaud/risq/risq.html`) proposé par Olivier Teytaud en 2006. Ce challenge s'est achevé par un workshop organisé à Paris en mai 2007.

En 2003, j'ai été membre du comité de pilotage de l'Action Spécifique (AS) du CNRS "Apprentissage et bioinformatique". J'ai animé dans ce cadre un groupe de travail portant sur l'apprentissage et le traitement de séquences : `http://www.loria.fr/~guermeur/GdT/`. J'ai également été membre de l'AS du CNRS intitulée "Machines à vecteurs support et méthodes à noyau".

Je suis membre de la "Fédération des Equipes de Recherche en Apprentissage" (FERA) : `http://www.lri.fr/~proml/wiki/index.php/Main_Page`.

De mars 2003 à novembre 2006, j'ai été le coordinateur du projet "Apprentissage automatique appliqué à la prédiction de la structure tertiaire des protéines" (GENOTO3D) sélectionné par

l'ACI "Masses de Données" en 2003 (Projet 2003-96). Ce projet, dont l'objectif était de prédire la structure tertiaire des protéines par des méthodes issues de l'apprentissage automatique, regroupait six équipes de six laboratoires :

1. Projet MODBIO du LORIA
2. Laboratoire de Bioinformatique et RMN Structurales (LBRS) de l'IBCP
3. Equipe "Bases de Données et Apprentissage Automatique" (BDAA) du LIF
4. Projet Symbiose de l'IRISA
5. Equipe "Méthodes et Algorithmes pour la Bioinformatique" (MAB) du LIRMM
6. Unité Mathématique, Informatique et Génome (MIG) de l'INRA, centre de Jouy-en-Josas

Les informations concernant ce projet sont disponibles sur son site web, dont l'adresse est la suivante : `http://www.loria.fr/~guermeur/ACIMD/`.

De 2005 à 2006, j'ai été membre du projet intitulé "Modélisation de la protéine FAK (Focal Adhesion Kinase) en vue de l'identification de molécules anti-métastases", dont le porteur était Bernard Maigret. Cette collaboration avec l'"équipe de Dynamique des Assemblages Membranaires" (eDAM) du laboratoire "Structure et Réactivité des Systèmes Moléculaires Complexes", UMR 7565 à Nancy, était une opération du thème "Bioinformatique et applications à la génomique" du PRST "Intelligence Logicielle".

D'octobre 2009 à octobre 2010, j'ai été le porteur du projet pilote P5-SVM-PROT2D du thème "Modélisation des Biomolécules et de leurs Interactions" (MBI) du CPER "Modélisations, Informations et Systèmes Numériques" (MISN). Ce projet s'est poursuivi jusqu'à la fin 2013, sous la forme de l'opération "Apprentissage statistique pour le traitement des problèmes de Discrimination sur les Séquences Biologiques" (ADiSBio) de l'axe "Coopération-Combinaison de Méthodes d'Analyse et de Fouille de Données" du thème MBI. Il avait pour objectif la mise à disposition progressive, depuis la plate-forme de bioinformatique du LORIA, des modules de la méthode de prédiction de la structure secondaire des protéines développée dans l'équipe ABC (voir la section 6.3).

4.2 Activités éditoriales

4.2.1 Organisation de sessions spéciales de conférences internationales

En 2000, Hélène Paugam-Moisy, André Elisseeff et moi avons organisé une session spéciale de la conférence "International Joint Conference on Neural Networks" (IJCNN) intitulée "Multi-Class Support Vector Machines".

En 2007, j'ai organisé la session spéciale de la conférence "Applied Stochastic Models and Data Analysis" (ASMDA 2007, `http://www.asmda.com/id7.html`) intitulée "Supervised Prediction with Neural Networks and SVMs" (`http://www.loria.fr/~guermeur/ASMDA_CFP.html`).

4.2.2 Comités de lecture

– Conférences internationales

1. "International Conference on Machine Learning" (ICML) 2008 (`http://icml2008.cs.helsinki.fi/`)
2. "Applied Stochastic Models and Data Analysis" (ASMDA) 2009 (`http://www.mii.lt/ASMDA%2D2009/`)
3. "International Conference on Stochastic Modeling Techniques and Data Analysis" (SMTDA) 2010 et 2012 (`http://smtda.net/`)
4. "European Conference on Artificial Intelligence" (ECAI) 2014 (`http://www.ecai2014.org/`)

5. "IAPR International Conference on Pattern Recognition in Bioinformatics" (PRIB) 2014 (http://prib2014.scilifelab.se/)
6. "Modelling, Computation and Optimization in Information Systems and Management Sciences" (MCO) 2015 (http://www.lita.univ-lorraine.fr/iccsama2015/MCO/index.php)

– Conférences nationales

1. "Conférence francophone d'Apprentissage" (CAp) depuis 2003
2. Congrès Francophone AFRIF-AFIA de "Reconnaissance des Formes et Intelligence Artificielle" (RFIA) 2004
3. "Maghrebian Conference on Information Technologies" 2008 (http://mcseai.simpa-usto.net/?page=committees)
4. "Journées Ouvertes en Biologie, Informatique et Mathématiques" (JOBIM) 2009 et 2010
5. "Journées de la société française de chémoinformatique" (SFCi) 2013 (http://sfci2013.loria.fr/)

– Ateliers de travail

1. "Workshop on Modelling and Optimization Techniques in Information Systems, Database Systems and Industrial Systems" (MOT-ACIIDS) 2013 (http://seminar.spaceutm.edu.my/aciids2013/specialsessions.html)
2. Atelier "Apprentissage et données Omiques" (AdO'13) de CAp'13 (https://sites.google.com/site/adpg2013/)
3. "International Workshop on Biological Knowledge Discovery and Data Mining" (BIOKDD) 2015 (http://www.dexa.org/biokdd2015)

– Divers

1. Challenge théorique PASCAL "Type I and type II errors for multiple simultaneous hypothesis testing" (http://www.lri.fr/~teytaud/risq/risq.html)
2. Numéro spécial de la "Revue des Nouvelles Technologies de l'Information" (RNTI) dédié aux journées thématiques "Apprentissage Artificiel & Fouille de Données" (AAFD) 2008 (http://www-lipn.univ-paris13.fr/A3/AAFD08/)

4.2.3 Relecture d'articles

Je suis relecteur régulier pour les principales revues d'apprentissage, parmi lesquelles le "Journal of Machine Learning Research" (JMLR), Machine Learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) et Neurocomputing. Mon travail d'évaluation concerne majoritairement des revues d'apprentissage, de statistique et de bioinformatique. Les principales sont à présent regroupées par domaine.

Apprentissage "Journal of Pattern Recognition Research" (JPRR), "Journal of Artificial Intelligence Research" (JAIR), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), IEEE Transactions on Systems, Man and Cybernetics (Part B) (SMCB), IEEE Transactions on Signal Processing, l'"International Journal of AI Tools" (IJAIT) et la revue RIA.

Statistique Statistics and Computing, Communications in Statistics - Theory and Methods, Journal of Computational and Graphical Statistics (JCGS) et "International Statistical Review" (ISR).

Bioinformatique - biochimie Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), "Internet Electronic Journal of Molecular Design" et "Process Biochemistry".

Autres "Discrete Applied Mathematics" (DAM), "Annals of Mathematics and Artificial Intelligence" (AMAI) et la RNTI.

J'ai également relu des communications soumises à de nombreuses conférences dont IJCAI, NIPS, COLT, ACML, ICANN, IJCNN, WABI, ECA, JOBIM et SFC.

4.3 Responsabilités collectives

De septembre 1999 à septembre 2003, j'ai été co-directeur des études du département "Services et Réseaux de Communication" (SRC) de l'IUT de Saint-Dié-des-Vosges, qui était une composante de l'UHP. J'ai appartenu aux différents jurys d'admission, de passage et de diplôme de cet IUT. De février 2000 à septembre 2003, j'ai également fait partie de la commission de choix. A ce titre, j'ai en particulier rapporté sur les dossiers de candidature aux postes de maîtres de conférences et d'ATER ouverts au recrutement au département. J'ai aussi participé aux travaux de la commission mixte constituée à partir de la commission de choix et de la commission de spécialistes des sections 7, 11, 12 et 71 de l'UHP.

De novembre 2005 à juin 2008, j'ai été membre suppléant élu de la commission de spécialistes de la section 27 de l'UHP. En juin 2007, j'ai été nommé membre titulaire de la commission de spécialistes de la section 27 de l'Université Paris 13.

En 2009, j'ai participé aux comités de sélection de deux postes de maîtres de conférences en section 27 : le poste de numéro 308 à l'UHP et celui de numéro 1225 à l'IUT d'Orsay.

De 2004 à 2006, j'ai été membre du comité des opérations du thème "Bioinformatique et applications à la génomique" du PRST "Intelligence Logicielle". J'ai poursuivi cette activité dans le cadre du thème "Modélisation des Biomolécules et de leurs Interactions" (MBI) du CPER "Modélisations, Informations et Systèmes Numériques" (MISN) qui s'étendait de 2007 à 2013.

Je suis le représentant des équipes ABC et CARTE à la Commission "Information et Edition Scientifique" (IES), anciennement "Commission Documentation" (ComDoc) puis "Information Scientifique et Technique" (IST), du LORIA.

4.4 Activités d'expertise

De 2003 à 2005, j'ai rédigé des rapports sur plusieurs projets soumis aux ACI "Masse de Données" et IMPBio. Depuis 2006, je suis expert pour l'ANR. Dans ce cadre, j'ai rédigé des rapports sur des projets soumis aux programmes "blanc", "Jeunes chercheuses - jeunes chercheurs" et "Masse de Données". En 2007, j'ai expertisé un projet soumis au programme "Plates-Formes Technologiques du Vivant" (PFTV) et en 2008, un projet soumis au programme "Domaines émergents : Nouveaux défis scientifiques et technologiques" (DEFIS). En 2012 et 2013, j'ai été membre du jury du prix de thèse de l'Association Française pour l'Intelligence Artificielle (AFIA) (`http://www.afia.asso.fr/tiki-index.php?page=Prix+de+Th%C3%A8se+en+Intelligence+Artificielle`).

4.5 Jurys de thèses et d'habilitations à diriger des recherches

Le 11 janvier 2006, j'ai participé en tant que membre invité au jury de thèse de Nicolas Sapay. Cette thèse de biologie de l'Université Lyon 1, dirigée par Gilbert Deléage et François Penin, est intitulée "Les peptides d'ancrages à l'interface membranaire, analyses structurales par RMN et dynamique moléculaire et développement d'une méthode de prédiction bioinformatique".

Le 12 décembre 2007, j'ai participé en tant qu'examinateur au jury de thèse de Christophe Magnan. Cette thèse d'informatique de l'Université de Provence - Aix-Marseille I, dirigée par François Denis et Cécile Capponi, est intitulée "Apprentissage à partir de données diversement étiquetées

pour l'étude du rôle de l'environnement local dans les interactions entre acides aminés".

Le 13 mars 2008, j'ai participé en tant qu'examinateur au jury de thèse d'Abderrahmane Boubezoul. Cette thèse d'informatique de l'Université Paul Cézanne - Aix-Marseille III, dirigée par Mustapha Ouladsine et Sébastien Paris, est intitulée "Système d'aide au diagnostic par apprentissage : application aux systèmes microélectroniques".

Le 10 octobre 2008, j'ai participé en tant qu'examinateur au jury de thèse de Nicolas Garnier. Cette thèse de biologie de l'Université Lyon 1, dirigée par Gilbert Deléage et Emmanuel Bettler, est intitulée "Mise en place d'un environnement bioinformatique d'évaluation et de prédiction de l'impact de mutations sur le phénotype de pathologies humaines".

Le 17 décembre 2008, j'ai participé en tant que rapporteur au jury de thèse de Bernhard Gschloessl. Cette thèse de biologie de l'Université Paris 6, dirigée par J. Mark Cock, est intitulée "Development of a method which predicts N-terminal target peptides and study of protein sorting in eukaryote genomes".

Le 9 juillet 2009, j'ai participé en tant que rapporteur au jury de thèse de Karina Zapién Arreola. Cette thèse d'informatique de l'INSA de Rouen, dirigée par Stéphane Canu, est intitulée "Algorithme de chemin de régularisation pour l'apprentissage statistique".

Le 5 janvier 2010, j'ai participé en tant qu'examinateur au jury de thèse de Cécile Bonnard. Cette thèse d'informatique de l'Université Montpellier II, dirigée par Olivier Gascuel et Nicolas Lartillot, est intitulée "Optimisation de potentiels statistiques pour un modèle d'évolution soumis à des contraintes structurales".

Le 15 décembre 2010, j'ai participé en tant que rapporteur au jury d'HDR de Nicolas Wicker. Cette HDR de l'Université de Strasbourg est intitulée "Données et méthodes aléatoires en biologie".

Le 28 octobre 2011, j'ai participé en tant que rapporteur au jury de thèse de Mamadou Thiao. Cette thèse de mathématiques appliquées de l'INSA de Rouen, dirigée par Tao Pham Dinh, est intitulée "Approches de la programmation DC et DCA en data mining".

Le 15 novembre 2011, j'ai participé en tant qu'examinateur au jury de thèse de Neeraj Kumar Singh. Cette thèse d'informatique de l'UHP, dirigée par Dominique Mery, est intitulée "Reliability and safety of critical device software systems".

Le 28 mars 2012, j'ai participé en tant que rapporteur au jury de thèse de Juliana Silva Bernardes. Cette thèse d'informatique de l'Université Pierre et Marie Curie et de l'"Universidade Federal do Rio de Janeiro", dirigée par Alessandra Carbone et Gerson Zaverucha, est intitulée "Combining evolution and machine learning for functional annotation and classification of remote homologous proteins".

Le 21 février 2013, j'ai participé en tant que rapporteur au jury de thèse de Mohammed Hindawi. Cette thèse d'informatique de l'INSA de Lyon dirigée par Alexandre Aussem, Khalid Benabdeslem et Jean-François Boulicaut, est intitulée "Sélection de Variables pour l'Analyse des Données Semi-Supervisées dans les Systèmes d'Information Décisionnels".

Le 7 novembre 2013, j'ai participé en tant que rapporteur au jury de thèse de Thanh-Nghi Doan. Cette thèse d'informatique de l'Université de Rennes 1, dirigée par François Poulet, est intitulée "Large Scale Support Vector Machines Algorithms for Visual Classification".

Le 19 mai 2014, j'ai participé en tant qu'examinateur au jury de thèse de Nguyen Manh Cuong. Cette thèse d'informatique de l'Université de Lorraine, dirigée par Le Thi Hoai An et Brieuc

Conan-Guez, est intitulée "La programmation DC et DCA pour certaines classes de problèmes en apprentissage et fouille de données".

Le 20 juin 2014, j'ai participé en tant qu'examinateur au jury d'HDR de Khalid Benabdeslem. Cette HDR d'informatique de l'Université Lyon 1 est intitulée "Contributions en apprentissage semi-supervisé : modélisation, classification et sélection".

Le 17 octobre 2014, j'ai participé en tant que rapporteur au jury de thèse de Rohit Babbar. Cette thèse d'informatique de l'Université de Grenoble, dirigée par Eric Gaussier et Massih-Reza Amini, est intitulée "Machine learning strategies for large-scale taxonomies".

Le 3 décembre 2014, j'ai participé en tant que rapporteur au jury de thèse de Nicolas Jung. Cette thèse de biologie des systèmes de l'Université de Strasbourg, dirigée par Seiamak Bahram et Myriam Maumy-Bertrand, est intitulée "Modélisation de phénomènes biologiques complexes : application à l'étude de la réponse antigénique de lymphocytes B sains et tumoraux".

4.6 Animation d'équipes de recherche

Depuis juillet 2006, j'anime l'activité scientifique de l'équipe ABC du LORIA, anciennement projet MODBIO de l'INRIA Lorraine. Le texte de notre projet, présenté à l'assemblée des responsables d'équipes (AREQ) du LORIA en septembre 2006, est disponible à l'adresse suivante : `http://modbio.loria.fr/download/abc2006.pdf`.

4.7 Organisation d'événements scientifiques

Samy Tindel et moi avons organisé la première journée scientifique de la Fédération Charles Hermite (`http://www.fr-hermite.univ-lorraine.fr/documents/2009609_1ere_journee_fch.html/index.html`). Son thème était l'apprentissage et elle s'est tenue au LORIA le 9 juin 2009.

5 Transfert technologique, relations industrielles et valorisation

5.1 Participation à des contrats de recherche

De septembre 2003 à mars 2007, j'ai participé au projet intitulé "Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques". Ce projet, une collaboration avec le laboratoire "Maturation des ARN et Enzymologie Moléculaire" (MAEM), UMR 7567 à Nancy, a été financé pendant 18 mois par le programme Décrypthon, partenariat entre l'AFM, le CNRS et IBM (`http://www.decrypthon.fr/`). De 2005 à 2006, il a également fait l'objet d'une opération du thème "Bioinformatique et applications à la génomique" du PRST "Intelligence Logicielle", opération dont j'étais co-responsable.

5.2 Participations à des travaux donnant lieu à des dépôts de brevets ou à des développements de logiciels

Deux logiciels de prédiction de la structure secondaire des protéines globulaires, HNN et MLRC, développés durant ma thèse, sont disponibles en ligne sur le serveur d'analyse de séquences protéiques NPS@ (`http://npsa-pbil.ibcp.fr/`) du Pôle Bio-Informatique Lyonnais (PBIL Lyon-Gerland, `http://pbil.univ-lyon1.fr/`), à l'adresse suivante : `http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html`.

Les travaux qu'Hélène Paugam-Moisy, André Elisseeff et moi avons effectués sur la théorie des SVM multi-classes m'ont conduit à développer un logiciel implémentant la M-SVM de Weston et Watkins de manière à supporter les problèmes de très grande taille. Ce logiciel, disponible depuis le site des "kernel machines" (`http://www.kernel-machines.org`, rubrique "Software") et depuis ma page web, a été déposé à l'Agence pour la Protection des Programmes (APP) le 18 avril 2005 sous le numéro IDDN.FR.001.170014.000.R.P.2005.000.10000. Une version dédiée au traitement des séquences biologiques est également disponible sur ma page web.

Le logiciel "AmphipaSeeK" mettant en œuvre la méthode de prédiction des ancrages membranaires interfaciaux introduite dans l'article de référence [JI06] dans ma liste de publications est disponible en ligne sur le serveur NPS@ du PBIL, à l'adresse suivante : `http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_amphipaseek.html`.

Le logiciel "HECTAR" mettant en œuvre la méthode de prédiction de la localisation cellulaire des protéines des hétérochontes introduite dans l'article de référence [JI08] dans ma liste de publications est disponible en ligne à l'adresse suivante : `http://www.sb-roscoff.fr/hectar/`. La variante nommée "HECTAR^{SEC}" (voir également [JI08]) peut être utilisée sur le même site.

J'ai programmé la SVM multi-classe à coût quadratique qu'Emmanuel Monfrini et moi avons introduite dans l'article de référence [JI10] dans ma liste de publications sous le nom de M-SVM^2. L'ensemble des codes correspondants a été enregistré à l'APP le 13 septembre 2010 sous le numéro `IDDN.FR.001.370001.000.S.P.2010.000.30000`. Le CNRS a également accordé sa diffusion sous licence CeCILL-B.

6 Résumé des travaux antérieurs et de leur impact scientifique et technologique

6.1 Introduction

Mon activité de recherche s'appuie sur deux axes principaux : la théorie statistique de l'apprentissage et la bioinformatique, plus particulièrement le traitement de séquences biologiques. Mon ambition est double. Elle consiste d'une part à faire progresser la théorie et la pratique de la discrimination à catégories multiples et d'autre part à effectuer des travaux dans lesquels la contribution originale relève conjointement de l'apprentissage et de la bioinformatique. De fait, l'observation qui fonde ma démarche est la suivante. La plupart des grands problèmes ouverts en bioinformatique soulèvent d'importantes questions théoriques et pratiques dans le domaine de la reconnaissance des formes et singulièrement en discrimination multi-classe. L'étude de ces questions se trouve encore souvent dans sa phase initiale. Réciproquement, les grandes avancées qui doivent avoir lieu en théorie statistique de l'apprentissage trouveront dans l'exploitation des données biologiques une pierre de touche sans égale.

6.2 Théorie statistique de l'apprentissage

La théorie statistique de l'apprentissage [83, 22, 84, 46] est un domaine de la statistique inférentielle dont les fondements ont été posés par V.N. Vapnik à la fin des années 60. Elle repose sur l'hypothèse que les éléments de l'espace de description et les valeurs qui peuvent leur être associées sont liés par une dépendance de nature probabiliste, caractérisée par une mesure de probabilité P qui est fixe mais inconnue. L'objet de cette théorie est la détermination des conditions sous lesquelles il est possible d'apprendre à partir de données empiriques, que l'on suppose par défaut avoir été obtenues par échantillonnage aléatoire simple suivant la loi P. L'apprentissage se conçoit principalement comme un problème de sélection de fonction ou sélection de modèle [46, 57]. Il s'agit de déterminer, dans une ou plusieurs classes de fonctions données, de cardinalités ordinairement infinies, une fonction permettant d'obtenir les meilleures performances possibles sur un problème

donné. Le problème en question peut relever de l'analyse discriminante, de l'approximation de fonctions (régression) ou de l'estimation de (la fonction de) densité.

Cette théorie étudie particulièrement deux principes inductifs. Le premier, nommé principe de minimisation empirique du risque, consiste à minimiser l'erreur en apprentissage. Dans le cas des petits échantillons, pour prendre en compte le fait que l'intervalle de confiance n'est pas négligeable devant le risque empirique, on substitue à ce principe celui de minimisation structurelle du risque (SRM) [83], consistant à minimiser une borne sur le risque (erreur en généralisation), borne souvent nommée "risque garanti" alors même qu'elle traduit un résultat de convergence presque sûrement uniforme. Ce dernier principe est en particulier celui qui a fourni leur cadre théorique initial aux problèmes d'apprentissage des machines à vecteurs support (SVM) [12, 20, 72]. En pratique, ces problèmes apparaissent de manière tout aussi naturelle en choisissant pour cadre la théorie de la régularisation "à la Tikhonov" [77].

Les SVM sont des modèles de l'apprentissage statistique initialement conçus pour le calcul des dichotomies. Elles se présentaient alors comme des extensions non linéaires de l'hyperplan de marge maximale [83]. Ces extensions découlent du remplacement dans les calculs du produit scalaire euclidien par une fonction de type positif ou noyau [9]. En pratique, cette modification revient à substituer à l'espace de description initial un espace de Hilbert à noyau reproduisant (RKHS) [5, 9]. Ainsi, les SVM constituent l'exemple le plus connu de machines à noyau [69, 70]. Le concept a ensuite été étendu afin de permettre l'approximation de fonctions et l'estimation de la fonction de densité. Le premier modèle de SVM multi-classe (M-SVM), permettant d'effectuer des tâches de discrimination à catégories multiples sans avoir recours à la mise en œuvre d'une méthode de décomposition, a été proposé indépendamment par plusieurs groupes d'auteurs (voir en particulier [84, 87, 14]). Depuis lors, de nouveaux modèles ont régulièrement été proposés (voir [34] pour un état de l'art).

Le champ d'application de la théorie statistique de l'apprentissage ne se limite pas aux machines à noyau, mais englobe également de nombreux autres outils de la statistique non paramétrique, comme les réseaux de neurones [4, 42] ou les systèmes de modélisation stochastique, tels les modèles de Markov cachés (HMM) [64]. En discrimination, le modèle formel qu'elle considère de manière privilégiée est celui du classifieur à vaste marge, dans lequel entrent aussi bien des machines à noyau que des réseaux neuronaux, comme le perceptron multi-couche (PMC), ou des méthodes de combinaison de classifieurs "faibles" telles le *boosting* [25, 26]. Au-delà de la marge géométrique au centre du problème d'apprentissage des SVM, la notion de marge varie avec les auteurs (voir par exemple [2, 56, 78, 31]), ce qui donne naissance à autant de résultats d'une grande utilité pratique. La convention utilisée dans toute la suite de ce document est la suivante : un système discriminant à marge désigne un classifieur prenant ses valeurs dans $\mathbb{R}^Q$, où $Q \in \mathbb{N} \setminus \{0, 1\}$ est le nombre de catégories. Pour chaque description, il fournit un score par catégorie, score qui peut correspondre ou non à une probabilité d'appartenance. La catégorie retenue est alors celle associée au score le plus élevé.

6.2.1 Etude des performances en généralisation des systèmes discriminants multi-classes à marge

La théorie statistique de la discrimination multi-classe est un domaine d'une importance considérable, y compris d'un point de vue pratique, qui n'a pourtant reçu de la communauté qu'une attention trop limitée. La raison en est que beaucoup considèrent suffisant d'établir une théorie pour le calcul des dichotomies, et de recourir ensuite à des méthodes de décomposition [49, 2, 66] si le nombre de catégories est supérieur à deux. Cependant, on peut démontrer que l'extension des bornes standard (généralisant le théorème de Glivenko-Cantelli [22, 84]) au cas multi-classe produit des majorants du risque de meilleure qualité que l'application combinée de bornes bi-classes. De plus, la généralisation des résultats établis pour les dichotomies est très rarement triviale. Il apparaît au contraire que les preuves de ces résultats s'appuient de manière essentielle sur les particularités

du cas bi-classe, comme la nature scalaire, et non vectorielle, des sorties du modèle étudié. Etablir la théorie statistique de la discrimination multi-classe soulève donc des problèmes difficiles, réellement originaux, qui ne peuvent être résolus de manière satisfaisante en se restreignant à l'arsenal mathématique utilisé jusqu'à présent pour les dichotomies.

Depuis mon recrutement au CNRS, en octobre 2003, l'essentiel de mon activité de recherche en apprentissage a eu pour but d'enrichir la théorie et la pratique de la discrimination multi-classe. J'ai tout d'abord proposé dans [30] une extension de la borne sur les performances en généralisation des classifieurs à grande marge introduite dans [7]. Ce résultat, qui étend également le théorème 4.1 de [84], fait intervenir comme mesure de capacité de la classe de fonctions considérée des nombres de couverture [50]. En s'inspirant de la démarche standard (voir par exemple [85, 22]), il est naturel de chercher à borner ces nombres en fonction de dimensions de Vapnik-Chervonenkis (VC) [85] généralisées. Cela nécessite de définir ces dimensions, d'établir les lemmes de Sauer-Shelah [68, 71] généralisés correspondants, et enfin de proposer des bornes sur les dimensions VC généralisées des modèles d'intérêt. Dans [31], j'ai apporté les solutions à ces problèmes, en mettant en évidence les propriétés intrinsèques du cas multi-classe. La théorie VC des systèmes discriminants multi-classes à grande marge est donc à présent établie. Elle repose sur les γ-Ψ-dimensions. Les familles de fonctions dont la capacité doit être évaluée (majorée) sont les images des familles de base (réalisables par les classifieurs) par différents opérateurs de marge. Le risque garanti obtenu (théorème 40) est perfectible, puisque le terme de contrôle correspondant décroît avec la taille m de l'échantillon d'apprentissage comme un $O\left(\ln(m) \cdot m^{-1/2}\right)$. Or, on sait que dans le cas bi-classe, le taux de décroissance optimal est un $O\left(m^{-1/2}\right)$ (voir par exemple le théorème 3.4 de [13]). Bien entendu, la dépendance à m ne doit pas varier avec le nombre de catégories.

6.2.2 Etude des performances en généralisation des M-SVM

En parallèle, avec Myriam Maumy et Frédéric Sur, j'ai suivi une autre voie pour borner les nombres de couverture dans le cas particulier des M-SVM. Nous nous sommes inspirés de l'étude dédiée aux machines à noyau (bi-classes) proposée dans [88, 89]. La démarche consiste à exprimer les nombres de couverture en fonction de nombres d'entropie [50, 17] d'opérateurs linéaires. Une fois cette transition effectuée, on a recours aux résultats sur les nombres d'entropie des opérateurs linéaires établis en analyse fonctionnelle [16, 17]. Ceci nous a permis de proposer une première méthode de sélection de modèle dédiée aux M-SVM [39]. Actuellement, elle n'est opérationnelle en pratique que lorsque l'espace de représentation (*feature space*) est de dimension finie. Ceci exclut des choix standard pour le noyau, comme celui du noyau à fonction de base radiale (RBF). Cette limitation sera levée lorsque nous aurons achevé l'extension du théorème de Maurey-Carl [17] consistant à utiliser des hypothèses plus générales tout en spécifiant les valeurs des constantes universelles, en particulier pour le cas dual.

Au début des années 2000, différents travaux ont mis en évidence l'intérêt de dériver des bornes sur le risque fondées sur une autre mesure de capacité : la complexité de Rademacher [8, 13]. Comme dans l'approche décrite au paragraphe précédent, il s'agit de tirer parti des propriétés de la classe fonctionnelle impliquée afin d'éviter de faire intervenir une dimension VC généralisée. L'intérêt de cette mesure réside dans le fait qu'elle se majore plus directement que les nombres de couverture pour un vaste ensemble de modèles, parmi lesquels des machines à noyau (voir par exemple [8]). Elle apparaît a priori particulièrement adaptée pour exploiter finement la propriété de reproduisance. Dans [33], j'ai étendu au cas des M-SVM la borne sur le risque des SVM fondée sur l'inégalité des différences bornées [58] et impliquant un terme de contrôle faisant intervenir une complexité de Rademacher. J'ai ensuite établi dans [36] qu'il n'était pas possible de majorer efficacement cette complexité par un calcul direct étendant le calcul du cas bi-classe [59] (on obtient ainsi une dépendance quadratique au nombre de catégories). Cela m'a conduit à introduire dans la dérivation de la borne une étape supplémentaire de "chaînage" [23, 24, 74]. Le chaînage est une technique bien connue en théorie des processus [81], qui est entre autres à la base de l'optimisation standard des bornes VC (voir par exemple le théorème 3.4 de [13]). Dans [36], elle intervient précisément pour

fournir une majoration de la complexité de Rademacher en fonction de nombres d'entropie. Pour borner les nombres d'entropie d'intérêt, différents résultats sont disponibles [80, 79]. La combinaison du chaînage et d'une méthode spécifique de majoration des nombres d'entropie produit la borne exposée dans [36], dont le taux de convergence est proche de l'optimal (en $O\left(\sqrt{\frac{\ln(m)}{m}}\right)$), tandis que la dépendance au nombre de catégories est meilleure que celle du théorème 8.1 de [59].

6.2.3 Sélection de modèle pour les M-SVM

Les bornes évoquées dans les deux sections précédentes, valables à taille d'échantillon finie, possèdent une utilité pratique. J'étudie actuellement leur efficacité en sélection de modèle, et plus particulièrement pour choisir la valeur du paramètre de marge douce C des différentes M-SVM proposées dans la littérature [34]. Naturellement, la méthode de référence en sélection de modèle est la validation croisée [73, 15], et plus particulièrement sa variante dite *leave-one-out*, qui présente l'avantage de fournir un estimateur de l'erreur en généralisation presque sans biais [55]. De nombreux travaux ont porté sur la majoration de l'erreur de validation croisée *leave-one-out* des SVM bi-classes. Parmi les bornes auxquelles ils ont donné naissance (voir [19] pour un état de l'art sur le sujet) la *span bound* de Chapelle et Vapnik [82] est considérée comme étant la plus fine. La plus utilisée est probablement la borne "rayon-marge" [84], qui représente un bon compromis entre qualité et temps de calcul. Plus précisément, elle s'avère presque aussi efficace que la *span bound* pour déterminer la valeur des hyper-paramètres, tandis que les calculs qu'elle nécessite sont moins lourds.

Dans [40], Emmanuel Monfrini et moi avons établi une borne "rayon-marge" dédiée à la M-SVM de Lee, Lin et Wahba [53] à marge dure. Nous avons également introduit la première M-SVM "à coût quadratique" : la M-SVM^2. Il s'agit d'une variante de la machine de Lee, Lin et Wahba à marge douce qui est à cette machine ce que la SVM "de norme 2" (ℓ_2-SVM) [20] est à la SVM "standard" (ℓ_1-SVM). Ainsi, notre borne "rayon-marge" généralisée s'applique à cette nouvelle machine, ce qui permet de l'utiliser pour déterminer la valeur du paramètre C. Rémi Bonidal et moi avons travaillé à la détermination d'algorithmes de parcours du chemin de régularisation dédiés aux SVM (bi-classes et multi-classes) "à coût quadratique". Il s'agit de réduire le temps de calcul nécessaire à la sélection de modèle, à l'instar de ce qui a déjà été fait pour deux machines utilisant la norme ℓ_1 sur le vecteur des variables d'écart, la SVM standard, dans [44], et la M-SVM de Lee, Lin et Wahba, dans [52]. Nous avons établi le fait que le passage de la norme ℓ_1 à un coût quadratique induit un comportement plus complexe, l'évolution des paramètres n'étant plus linéaire par morceaux. S'en suit la nécessité de trouver des solutions efficaces pour que ce parcours s'effectue en un temps acceptable, sans poser de problème de nature numérique. Celle que nous avons privilégiée dans le cas bi-classe (ℓ_2-SVM), qui rejoint un algorithme proposé par Olivier Chapelle dans [18], présente l'avantage de permettre le calcul de la *span bound* pour un coût additionnel très faible. Notre algorithme intégrant parcours du chemin de régularisation et sélection de modèle pour cette SVM est décrit dans [11]. Rémi a également produit un algorithme de parcours du chemin de régularisation dédié à la M-SVM^2. La *span bound* correspondante a été obtenue dans le cadre d'une collaboration avec Liva Ralaivola, du LIF, à l'Université de Provence. La conclusion à laquelle nous sommes parvenus est que pour cette machine et l'algorithme de parcours du chemin précité, le critère de sélection de modèle le plus approprié (réalisant le meilleur compromis entre qualité du résultat et temps de calcul) est cette fois la borne "rayon-marge" correspondante. Ici encore apparaît donc une différence instructive entre le calcul des dichotomies et celui des polytomies. En pratique, le premier critère à considérer pour choisir entre minimiser cette borne le long du chemin ou au moyen d'une simple descente en gradient est le rang de la matrice de Gram.

6.2.4 Mise en œuvre des M-SVM

L'encadrement de la thèse de Rémi Bonidal, évoqué dans la section précédente, comme celui de la thèse d'Hafida Chouarfia, m'ont conduit à partir de 2009 à m'intéresser plus avant aux aspects

pratiques de la mise en œuvre des M-SVM. De ce point de vue, un champ d'investigation qui restait largement ouvert est la combinaison de M-SVM, en lien avec l'estimation des probabilités a posteriori des catégories. La solution la plus populaire pour dériver de telles estimations à partir des sorties de SVM est le couplage par paire (*pairwise coupling*) [45]. Elle s'applique aux sorties post-traitées de SVM bi-classes, dans le cadre de la mise en œuvre de la méthode de décomposition "un contre un". Le post-traitement standard, introduit par Platt [62] et amélioré dans [54], consiste en une régression logistique (application d'une fonction sigmoïde paramétrée). Ma contribution initiale à la combinaison de M-SVM a pris la forme de l'évaluation de "méthodes d'ensemble linéaires" (LEM) [35]. Cette étude rejoignait sur certains points mes travaux de thèse [28, 37], avec une amélioration notable : l'introduction des γ-Ψ-dimensions permet de caractériser précisément la complexité en échantillon des méthodes d'ensemble considérées. Ce point est important, dans la mesure où le sur-apprentissage (*over-fitting*) constitue la première limite rencontrée en combinaison de modèles. Une observation réalisée dans le cadre de ces recherches est que l'utilisation de la régression logistique polytomique [47] pour post-traiter les sorties des M-SVM de manière à obtenir des estimations des probabilités a posteriori des catégories présente un inconvénient : son emploi peut faire décroître significativement le taux de reconnaissance. Cette observation m'a conduit à infléchir les travaux d'Hafida Chouarfia vers l'estimation de ces probabilités à partir des sorties d'un ensemble de M-SVM. Dans le cadre plus vaste de la combinaison de classifieurs à marge (M-SVM, réseaux de neurones...), Fabienne Thomarat, Rémi Bonidal et moi avons développé un modèle hiérarchique et modulaire nommé MSVMpred [10] visant à optimiser conjointement le taux de reconnaissance et la qualité des estimations des probabilités a posteriori des catégories. Le résultat majeur de ces travaux est que les performances de MSVMpred au sens des deux critères précités sont supérieures à celles du couplage par paire. Le gain s'avère statistiquement significatif pour un grand nombre de jeux de données parmi les plus utilisés dans le domaine de l'apprentissage automatique.

A la fin de l'année 2010, j'ai introduit dans [34] le premier modèle générique de M-SVM. Il plonge l'ensemble des machines de ce type publiées à ce jour dans un cadre théorique unificateur relevant à la fois de la théorie des RKHS de fonctions à valeurs vectorielles [86] et de la régularisation "à la Tikhonov". Son intérêt est double. D'une part, il permet la conception de nouvelles machines exhibant des propriétés spécifiques. Il a ainsi donné naissance à la classe des M-SVM "à coût quadratique". D'autre part, il permet une analyse globale des propriétés statistiques des M-SVM, au premier rang desquelles se trouve la consistance [91, 75]. Fabien Lauer et moi l'avons mis en œuvre dans une librairie C nommée MSVMpack [51].

6.3 Bioinformatique

En bioinformatique, je m'intéresse essentiellement aux problèmes de traitement de séquences, plus particulièrement ceux relatifs au repliement des protéines. Mes travaux ont donc en premier lieu relevé de la biologie structurale prédictive.

6.3.1 Prédiction de la structure secondaire des protéines globulaires

La prédiction de la structure secondaire des protéines globulaires est l'un des principaux problèmes ouverts en biologie structurale. Considérée du point de vue de la reconnaissance des formes, elle se formule comme un problème de discrimination à trois catégories consistant à affecter à chaque résidu d'une chaîne protéique sont état conformationnel : hélice α, brin β ou apériodique (coil). Elle représente le thème de recherche fédérateur de l'équipe ABC en bioinformatique. Nous l'étudions régulièrement, aussi bien dans le cadre de collaborations avec des biologistes et des bioinformaticiens qu'en interne. Elle constitue en particulier l'un des deux volets du sujet de thèse d'Hafida Chouarfia, l'autre volet étant la prédiction de la structure tertiaire. Depuis l'époque de ma thèse, je l'utilise fréquemment comme preuve de concept, afin d'obtenir une première évaluation des idées que je développe en reconnaissance des formes, et singulièrement dans le domaine

du traitement de séquences (voir par exemple [29, 41, 38, 32]). Je propose pour cette prédiction une architecture hybride, modulaire et hiérarchique. Inspirée par les célèbres travaux en parole associant réseaux de neurones et HMM (voir en particulier [60]), elle fait jouer un rôle central à la combinaison de modèles. Mon activité de fond en bioinformatique s'organise principalement autour du développement de modules de cette architecture, modules qui doivent pouvoir être utilisés pour d'autres tâches de traitement de séquences biologiques.

En 2008, j'ai effectué une étude comparative des performances en prédiction de la structure secondaire de l'ensemble des modèles de M-SVM publiés à l'époque [32]. Ces machines étaient munies du noyau dédié introduit dans [38]. Les résultats se sont révélés doublement positifs. D'une part, ils ont établi la supériorité de ces modèles sur le PMC, l'unité de base d'une majorité des méthodes de prédiction parmi les plus performantes à l'époque [21, 61], ainsi que sur les méthodes de décomposition impliquant des SVM bi-classes [48]. D'autre part, ils ont mis en évidence le fait que les M-SVM exhibent des comportements significativement différents, ce qui constitue un facteur favorable dans l'optique de leur combinaison.

Ceci nous a incités, Fabienne Thomarat et moi, à développer à partir de 2010 une instance de MSVMpred dédiée à la prédiction de la structure secondaire : MSVMpred2 [43]. Dans cette variante, les classifieurs de base, effectuant la prédiction dite "séquence-structure", sont des M-SVM et des réseaux de neurones récurrents, les BRNN [6, 63]. Ces classifieurs exploitent en entrée le contenu d'une fenêtre d'analyse glissant sur les lignes d'une *position-specific scoring matrix* (PSSM) produite par PSI-BLAST [3]. Leur combinaison est effectuée au moyen de différents classifieurs produisant la prédiction dite "structure-structure". Ces classifieurs, estimant tous les probabilités a posteriori des catégories, sont choisis de manière à couvrir un vaste spectre en termes de capacité. Enfin, la prédiction globale est obtenue par combinaison convexe des sorties des classifieurs "structure-structure". La combinaison convexe permet d'adapter de manière souple la complexité de la cascade de traitements, en fonction par exemple de la taille de la base d'apprentissage disponible. Après la mise au point de ces traitements de bas niveau, le travail s'est poursuivi en 2011, en collaboration avec Fabien Lauer, par l'interfaçage avec un traitement de plus haut niveau, effectué par un modèle génératif [76]. Ce modèle est une variante du "HMM inhomogène" (IHMM) proposé dans [65]. Le post-traitement des sorties de MSVMpred2 permet de prendre en compte dans la prédiction des dépendances à plus long terme, de même que des règles syntaxiques régissant l'ordonnancement des éléments conformationnels. Ceci se traduit par une amélioration statistiquement significative de l'une des mesures de qualité de la prédiction parmi les plus importantes : le coefficient Sov [90]. Actuellement, le taux de reconnaissance de notre méthode de prédiction hybride approche les 82%, pour une valeur du coefficient Sov égale à 80%. Ces performances constituent l'état de l'art. L'ensemble de ces recherches se concrétise par la production d'une méthode de prédiction entièrement développée par ABC, méthode qui est progressivement mise à la disposition de la communauté scientifique depuis la plate-forme de bioinformatique du LORIA.

6.3.2 Autres collaborations avec des biologistes et des bioinformaticiens

De 2005 à 2006, en collaboration avec Nicolas Sapay, étudiant en thèse à l'IBCP, à Lyon, sous la direction conjointe de Gilbert Deléage et François Penin, j'ai contribué à mettre au point la méthode "AmphipaSeeK" de prédiction des ancrages membranaires interfaciaux dans les protéines membranaires monotopiques [67]. Cette étude abordait un problème de biologie structurale particulièrement difficile, pour lequel peu de données étaient jusqu'alors disponibles. Du point de vue de l'apprentissage, son intérêt principal était de permettre une évaluation dans un contexte différent du noyau évoqué dans la section précédente, ce qui nous a en particulier conduits à mener une étude comparative portant sur les matrices de substitution. Nous avons également introduit dans la pratique l'emploi de SVM *topology-to-topology*, qui a depuis été repris par d'autres équipes.

Ma collaboration la plus récente avec un laboratoire de biologie est une collaboration avec Bernhard Gschloessl, de la station biologique de Roscoff, étudiant en thèse sous la direction de

J. Mark Cock. Elle a débuté à la fin de 2006 et portait sur la prédiction de la localisation cellulaire des protéines des hétérochontes. La méthode que nous avons développée, nommée "HEterokont subCellular TARgeting" (HECTAR) [27], est une méthode hiérarchique fondée sur une structure d'arbre. Les systèmes discriminants placés sur les nœuds de l'arbre sont des SVM et des M-SVM. Les résultats expérimentaux obtenus sont meilleurs que ceux fournis par les méthodes constituant l'état de l'art, la différence étant statistiquement significative. Les principales difficultés relevant de l'apprentissage qu'a soulevées la mise au point d'HECTAR sont de deux ordres. D'une part, le choix des prédicteurs à chaque nœud de l'arbre a dû faire l'objet d'investigations systématiques, afin d'accorder l'expertise du biologiste avec les capacités des systèmes discriminants. D'autre part, il a été nécessaire de développer des solutions optimisées pour mettre en œuvre la validation croisée sur cette structure hiérarchique de manière à contrôler les performances en généralisation. En ce sens, ce travail présentait d'importantes similitudes avec celui décrit au paragraphe précédent.

En septembre 2010, j'ai débuté une collaboration avec Faiza Abdat et Walter Blondel, du CRAN, à Nancy. Elle porte sur le diagnostic des étapes précédant le cancer de la peau chez la souris. La tâche à accomplir est un problème de discrimination à quatre catégories. Ici encore, les systèmes discriminants employés sont des M-SVM. Les données de base sont les spectres d'autofluorescence et de réflectance diffuse. La principale originalité de notre contribution réside dans le choix des transformations appliquées à ces spectres pour obtenir les prédicteurs fournis aux M-SVM [1].

Références

[1] F. Abdat, M. Amouroux, Y. Guermeur, and W. Blondel. Hybrid feature selection and SVM-based classification for mouse skin precancerous stages diagnosis from bimodal spectroscopy. *Optics Express*, 20(1) :228–244, 2012.

[2] E.L. Allwein, R.E. Schapire, and Y. Singer. Reducing multiclass to binary : A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1 :113–141, 2000.

[3] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Research*, 25(17) :3389–3402, 1997.

[4] M. Anthony and P.L. Bartlett. *Neural Network Learning : Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

[5] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3) :337–404, 1950.

[6] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11) :937–946, 1999.

[7] P.L. Bartlett. The sample complexity of pattern classification with neural networks : The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2) :525–536, 1998.

[8] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities : Risk bounds and structural results. *Journal of Machine Learning Research*, 3 :463–482, 2002.

[9] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, 2004.

[10] R. Bonidal, F. Thomarat, and Y. Guermeur. Estimating the class posterior probabilities in biological sequence segmentation. In *SMTDA'12*, 2012.

[11] R. Bonidal, S. Tindel, and Y. Guermeur. Model selection for the ℓ_2-SVM by following the regularization path. *Transactions on Computational Collective Intelligence*, XIII :83–112, 2014.

[12] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

[13] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : A survey of some recent advances. *ESAIM : Probability and Statistics*, 9 :323–375, 2005.

[14] E.J. Bredensteiner and K.P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12(1/3) :53–79, 1999.

[15] P. Burman. A comparative study of ordinary cross-validation, ν-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3) :503–514, 1989.

[16] B. Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Annales de l'institut Fourier*, 35(3) :79–118, 1985.

[17] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

[18] O. Chapelle. Training a support vector machine in the primal. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large-Scale Kernel Machines*, chapter 2, pages 29–50. The MIT Press, Cambridge, MA, 2007.

[19] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1) :131–159, 2002.

[20] C. Cortes and V.N. Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, 1995.

[21] J.A. Cuff and G.J. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins : Structure, Function, and Genetics*, 40(3) :502–511, 2000.

[22] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[23] R.M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3) :290–330, 1967.

[24] R.M. Dudley. A course on empirical processes. In P.L. Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XII - 1982*, volume 1097 of *Lecture Notes in Mathematics*, pages 1–142. Springer-Verlag, 1984.

[25] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2) :256–285, 1995.

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

[27] B. Gschloessl, Y. Guermeur, and J.M. Cock. HECTAR : A method to predict subcellular targeting in heterokonts. *BMC Bioinformatics*, 9(393), 2008.

[28] Y. Guermeur. *Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines*. Thèse de doctorat, Université Paris 6, 1997.

[29] Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, 5(2) :168–179, 2002.

[30] Y. Guermeur. Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions. Technical Report RR-5314, INRIA, 2004. (révisé en 2006).

[31] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8 :2551–2594, 2007.

[32] Y. Guermeur. Etude comparée des performances de SVM multi-classes en prédiction de la structure secondaire des protéines. *Revue des Nouvelles Technologies de l'Information*, A-3 :21–48, 2009.

[33] Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics - Theory and Methods*, 39(3) :543–557, 2010.

[34] Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6) :555–577, 2012.

[35] Y. Guermeur. Combining multi-class SVMs with linear ensemble methods that estimate the class posterior probabilities. *Communications in Statistics - Theory and Methods*, 42(16) :3011–3030, 2013.

[36] Y. Guermeur. Guaranteed risk for large margin multi-category classifiers. 2014. (soumis).

[37] Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deléage. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15(5) :413–421, 1999.

[38] Y. Guermeur, A. Lifchitz, and R. Vert. A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, chapter 9, pages 193–206. The MIT Press, Cambridge, MA, 2004.

[39] Y. Guermeur, M. Maumy, and F. Sur. Model selection for multi-class SVMs. In *ASMDA'05*, pages 507–517, 2005.

[40] Y. Guermeur and E. Monfrini. A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica*, 22(1) :73–96, 2011.

[41] Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy, and P. Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, 56 :305–327, 2004.

[42] Y. Guermeur and O. Teytaud. Estimation et contrôle des performances en généralisation des réseaux de neurones. In Y. Bennani, editor, *Apprentissage Connexionniste*, chapter 10, pages 279–342. Hermès, 2006.

[43] Y. Guermeur and F. Thomarat. Estimating the class posterior probabilities in protein secondary structure prediction. In *PRIB'11*, pages 260–271, 2011.

[44] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5 :1391–1415, 2004.

[45] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2) :451–471, 1998.

[46] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[47] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, London, 1989.

[48] S. Hua and Z. Sun. A novel method of protein secondary structure prediction with high segment overlap measure : Support vector machine approach. *Journal of Molecular Biology*, 308 :397–407, 2001.

[49] S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited : A stepwise procedure for building and training a neural network. In F. Fogelman-Soulié and J. Hérault, editors, *Neurocomputing : Algorithms, Architectures and Applications*, volume F68 of *NATO ASI Series*, pages 41–50. Springer-Verlag, 1990.

[50] A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations, series 2*, 17 :277–364, 1961.

[51] F. Lauer and Y. Guermeur. MSVMpack : a multi-class support vector machine package. *Journal of Machine Learning Research*, 12 :2293–2296, 2011.

[52] Y. Lee and Z. Cui. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 16(2) :391–409, 2006.

[53] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines : Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465) :67–81, 2004.

[54] H.-T. Lin, C.-J. Lin, and R.C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3) :267–276, 2007.

[55] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3, 1969. (en russe).

[56] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX(2) :245–303, 2000.

[57] P. Massart. *Concentration Inequalities and Model Selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Springer-Verlag, Berlin, 2007.

[58] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[59] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, 2012.

[60] N. Morgan, H. Bourlard, S. Renals, M. Cohen, and H. Franco. Hybrid neural network/hidden Markov model systems for continuous speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4) :899–916, 1993.

[61] T.N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G.P. Gippert, and O. Lund. Prediction of Protein Secondary Structure at 80% Accuracy. *Proteins : Structure, Function, and Genetics*, 41(1) :17–20, 2000.

[62] J.C. Platt. Probabilities for SV machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, chapter 5, pages 61–73. The MIT Press, Cambridge, MA, 2000.

[63] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins : Structure, Function, and Genetics*, 47(2) :228–235, 2002.

[64] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[65] P. Ramesh and J.G. Wilpon. Modeling state durations in hidden Markov models for automatic speech recognition. In *ICASSP-92*, volume I, pages 381–384, 1992.

[66] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5 :101–141, 2004.

[67] N. Sapay, Y. Guermeur, and G. Deléage. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. *BMC Bioinformatics*, 7(255), 2006.

[68] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1) :145–147, 1972.

[69] B. Schölkopf and A.J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.

[70] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

[71] S. Shelah. A combinatorial problem ; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1) :247–261, 1972.

[72] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

[73] M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1) :29–35, 1977.

[74] M. Talagrand. *The Generic Chaining : Upper and Lower Bounds of Stochastic Processes*. Springer-Verlag, Berlin, 2005.

[75] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8 :1007–1025, 2007.

[76] F. Thomarat, F. Lauer, and Y. Guermeur. Cascading discriminant and generative models for protein secondary structure prediction. In *PRIB'12*, pages 166–177, 2012.

[77] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. V.H. Winston & Sons, Washington, D.C., 1977.

[78] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1) :135–166, 2004.

[79] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, 2000.

[80] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

[81] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes, With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[82] V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9) :2013–2036, 2000.

[83] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[84] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.

[85] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2) :264–280, 1971.

[86] G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *SFI Studies in the Sciences of Complexity*, pages 95–112. 1992.

[87] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.

[88] R.C. Williamson, A.J. Smola, and B. Schölkopf. Entropy numbers of linear function classes. In *COLT'00*, pages 309–319, 2000.

[89] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines *via* entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6) :2516–2532, 2001.

[90] A. Zemla, Č. Venclovas, K. Fidelis, and B. Rost. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins : Structure, Function, and Genetics*, 34(2) :220–223, 1999.

[91] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5 :1225–1251, 2004.