Projet scientifique de l'équipe ABC

Apprentissage et Biologie Computationnelle

26 septembre 2006

Table des matières

1 Contexte **3**

2 Composition de l'équipe **3**

3 Objectifs généraux **3**

4 Fondements scientifiques **4**

5 Axes de recherche **6**
5.1 Apprentissage automatique 6
5.1.1 Bornes sur les performances en généralisation des systèmes discriminants multi-classes 7
5.1.2 Théorie et pratique des machines à noyau 9
5.1.3 Apprentissage non supervisé 9
5.2 Biologie computationnelle 10
5.2.1 Ingénierie du noyau 10
5.2.2 Développement d'architectures hybrides, intégrant systèmes discriminants et génératifs 10

6 Domaines d'application **11**

7 Logiciels **11**
7.1 Apprentissage automatique 11
7.2 Bioinformatique 11

8 Participation à des projets au niveau local, national et international **12**
8.1 Apprentissage automatique 12
8.2 Bioinformatique 13

9 Autres collaborations **13**

10 Bibliographie **14**

1 Contexte

L'équipe MODBIO a été constituée par Alexander Bockmayr, Professeur à l'Université Nancy 1, le premier janvier 2001. Il s'agissait alors, à notre connaissance, de la première équipe de bioinformatique du LORIA et de l'INRIA Lorraine. Après le départ de son responsable scientifique, à la fin de l'année 2004, l'équipe a poursuivi ses activités, en profitant en particulier, à la rentrée 2005, du renfort constitué par l'arrivée d'un nouveau membre permanent, Fabienne Thomarat. Le projet INRIA MODBIO ayant été arrêté le premier juillet 2006, elle postule au statut d'équipe LORIA et présente son projet de recherche sous l'intitulé "Apprentissage et Biologie Computationnelle" (ABC). L'utilisation de "biologie computationnelle" à la place de "bioinformatique" dans le nouveau nom de l'équipe et la définition des axes de recherche souligne une volonté de conserver un lien fort avec notre projet scientifique initial, dans lequel la modélisation jouait un rôle central.

2 Composition de l'équipe

Personnel CNRS et Université

Yann Guermeur [CR]
Fabienne Thomarat [Maître de conférences, INPL-Ecole des Mines de Nancy]

Stagiaires post-doctoraux

Emmanuel Monfrini [Université Henri Poincaré (UHP), Nancy 1]

Chercheurs doctorants

Yannick Darcy [Allocataire MENRT, UHP]

Collaborateurs extérieurs

Frédéric Bertrand [ATER, Université Louis Pasteur (ULP), Strasbourg 1]
Myriam Maumy-Bertrand [Maître de conférences, ULP]
Alexander Bockmayr [Professeur, "Freie Universität Berlin"]

3 Objectifs généraux

L'objectif de notre équipe est double :

- effectuer des recherches dans le domaine de la classification et de manière plus spécifique développer la composante à la fois la plus utile et la moins avancée de sa théorie, la théorie statistique de la discrimination multi-classe,
- utiliser les résultats obtenus pour mettre au point des méthodes à noyau dédiées à la reconnaissance des formes et appliquées de manière privilégiée à des tâches relevant de la biologie moléculaire.

4 Fondements scientifiques

La théorie statistique de l'apprentissage [Vap82,DGL96,Vap98] est un domaine de la statistique inférentielle dont les fondements ont été posés par V.N. Vapnik à la fin des années 60.

Sous l'hypothèse que les éléments de l'espace de description et les valeurs qui peuvent leur être associées sont liés par une dépendance probabiliste fixe mais inconnue, l'objet de cette théorie est de déterminer les conditions sous lesquelles il est possible d'apprendre à partir de données empiriques (obtenues par échantillonnage aléatoire simple suivant la loi jointe).

L'apprentissage se conçoit comme un problème de sélection de fonction ou sélection de modèle. 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. L'apprentissage se reformule ainsi comme un problème d'optimisation (précisément, un problème de M-estimation [vdG00]) consistant à minimiser une fonctionnelle nommée risque. Les domaines qu'il traite sont l'analyse discriminante, l'approximation de fonctions (régression) et l'estimation de la fonction de densité.

Cette théorie étudie particulièrement deux principes inductifs.

- Le premier, nommé principe de minimisation empirique du risque (ERM), consiste à minimiser l'erreur en apprentissage.
- Dans le cas des "petits" échantillons, on substitue à ce principe celui de minimisation structurelle du risque (SRM), consistant à minimiser une borne sur l'espérance du risque (erreur en généralisation) souvent nommée risque garanti. Ce dernier principe est en particulier mis en œuvre dans les algorithmes d'apprentissage des machines à vecteurs support (SVM) [BGV92,CV95], qui obtiennent actuellement les meilleures performances sur de nombreuses tâches relevant des principaux domaines de la reconnaissance des formes.

Les SVM constituent l'exemple le plus connu de machines à noyau [SS02]. Ces machines réalisent des familles de fonctions construites autour d'un noyau reproduisant [Aro50,Wah90, Wah99,BTA04]. Au-delà des SVM, la théorie de l'apprentissage s'intéresse particulièrement à leur

[Vap82] V. Vapnik, *Estimation of Dependences Based on Empirical Data.*, Springer-Verlag, N.Y, 1982.

[DGL96] L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[Vap98] V. Vapnik, *Statistical learning theory.*, John Wiley & Sons, Inc., N.Y., 1998.

[vdG00] S. van der Geer, *Empirical Processes in M-estimation*, Cambridge University Press, 2000.

[BGV92] B. Boser, I. Guyon, V. Vapnik, "A training algorithm for optimal margin classifiers", *in : COLT'92*, p. 144–152, 1992.

[CV95] C. Cortes, V. Vapnik, "Support-Vector Networks", *Machine Learning 20*, 1995, p. 273–297.

[SS02] B. Schölkopf, A. Smola, *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*, The MIT Press, 2002.

[Aro50] N. Aronszajn, "Theory of reproducing kernels", *Trans. Amer. Math. Soc. 68*, 1950, p. 337–404.

[Wah90] G. Wahba, "Spline Models for Observational Data", *in : SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics, 59*, 1990.

[Wah99] G. Wahba, "Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV", *in : Advances in Kernel Methods, Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola (editors), The MIT Press, 1999, p. 69–88.

[BTA04] A. Berlinet, C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Sta-*

mise en œuvre, dont elle permet d'analyser finement les propriétés statistiques. Un lien fort existe ainsi entre leur théorie et celle de la régularisation [TA77,BM91], qui se traduit entre autres dans la pratique par l'utilisation de termes de pénalisation identiques. Ce lien est par exemple mis en évidence dans [RC01].

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, comme les réseaux de neurones [Bis95,AB99] ou les systèmes de modélisation stochastique, tels les modèles de Markov cachés (HMM) [Rab89] ou les récents modèles de Markov couple (PMM) [Pie03] et modèles de Markov triplet (TMM) [Pie02]. En discrimination (cas du calcul des dichotomies), 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, ou des méthodes de combinaison de classifieurs faibles telles le boosting [Fre95,FS97]. Au-delà de la marge géométrique au centre de l'algorithme des SVM, la notion de marge, bi-classe (voir par exemple [Mas00,Tsy04]), ou multi-classe ([5],[ASS00]) varie avec les auteurs, ce qui donne naissance à autant de résultats d'une grande utilité pratique.

tistics, Kluwer Academic Publishers, 2004.

[TA77] A. Tikhonov, V. Arsenin, *Solutions of Ill-posed Problems*, W.H. Wiston and sons, Washington D.C., 1977.

[BM91] V. Badeva, V. Morozov, *Problèmes incorrectement posés*, Masson, 1991.

[RC01] A. Rakotomamonjy, S. Canu, "Frame, Reproducing Kernel, Regularization and Learning", *research report number 02-001*, P.S.I. INSA de Rouen, 2001.

[Bis95] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[AB99] M. Anthony, P. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, 1999.

[Rab89] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE 77*, 2, 1989, p. 257–286.

[Pie03] W. Pieczynski, "Pairwise Markov chains", *IEEE Trans. on PAMI 25*, 5, 2003, p. 634–639.

[Pie02] W. Pieczynski, "Chaînes de Markov triplet", *Comptes Rendus de l'Académie des Sciences - Mathématiques, Série I 335*, 3, 2002, p. 275–278.

[Fre95] Y. Freund, "Boosting a weak learning algorithm by majority", *Information and Computation 121*, 1995, p. 256–285.

[FS97] Y. Freund, R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences 55*, 1, 1997, p. 119–139.

[Mas00] P. Massart, "Some Applications of Concentration Inequalities to Statistics", *Annales de la Faculté des Sciences de Toulouse IX*, 2, 2000, p. 245–303.

[Tsy04] A. Tsybakov, "Optimal aggregation of classifiers in statistical learning", *Annals of Statistics 32*, 1, 2004.

[ASS00] E. Allwein, R. Schapire, Y. Singer, "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers", *Journal of Machine Learning Research 1*, 2000, p. 113–141.

5 Axes de recherche

En dépit de son développement rapide, s'appuyant sur des apports statistiques toujours nouveaux, issus des inégalités de concentration [Lug04] ou de la théorie des processus empiriques [SW86,Pol90,vdVW96,vdG00], la théorie statistique de l'apprentissage recèle encore de nos jours un domaine trop peu exploré, celui de la discrimination multi-classe. Actuellement, les principaux travaux dans le domaine font en fait intervenir des modèles bi-classes, que ce soit dans le cadre de méthodes de décomposition "un-contre-tous" [SBV95,Vap95,RK04], "un-contre-un" [MA98, WW98], ou de méthodes fondées sur l'utilisation de codes correcteurs d'erreurs (ECOC) [DB95, ASS00]. Ceci est particulièrement regrettable, pour deux raisons.

- D'une part, l'utilité d'une théorie statistique de la discrimination multi-classe apparaît de manière immédiate.
- D'autre part, celle-ci ne peut être obtenue comme une extension triviale de celle du calcul des dichotomies (on pourra sur ce point se rapporter à [7, 19]).

La composante apprentissage de notre projet de recherche vise en premier lieu à combler ce manque.

5.1 Apprentissage automatique

Notre activité en apprentissage s'organisera autour de trois thèmes principaux, le premier d'entre-eux étant les bornes sur le risque. Ce thème est celui sur lequel nous sommes actuellement le plus avancés, et il devrait servir, au moins dans un premier temps, à fédérer l'équipe.

[Lug04] G. Lugosi, "Concentration-of-measure inequalities", Lecture notes, Summer School on Machine Learning at the Australian National University, Canberra, 2004.

[SW86] G. Shorack, J. Wellner, *Empirical Processes with Applications to Statistics*, Wiley, New York, 1986.

[Pol90] D. Pollard, "Empirical Processes: Theory and Applications", *in : NFS-CBMS Regional Conference Series in Probability and Statistics, 2*, Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[vdVW96] A. van der Vaart, J. Wellner, *Weak Convergence and Empirical Processes, With Applications to Statistics, Springer Series in Statistics*, Springer-Verlag New York, Inc., 1996.

[vdG00] S. van der Geer, *Empirical Processes in M-estimation*, Cambridge University Press, 2000.

[SBV95] B. Schölkopf, C. Burges, V. Vapnik, "Extracting support data for a given task", *in : ICKDDM'95*, p. 252–257, 1995.

[Vap95] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, N.Y., 1995.

[RK04] R. Rifkin, A. Klautau, "In Defense of One-Vs-All Classification", *Journal of Machine Learning Research 5*, 2004, p. 101–141.

[MA98] E. Mayoraz, E. Alpaydin, "Support Vector Machines for Multi-Class Classification", *research report number 98-06*, IDIAP, 1998.

[WW98] J. Weston, C. Watkins, "Multi-class Support Vector Machines", *research report number CSD-TR-98-04*, Royal Holloway, University of London, Department of Computer Science, 1998.

[DB95] T. Dietterich, G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes", *Journal of Artificial Intelligence Research 2*, 1995, p. 263–286.

[ASS00] E. Allwein, R. Schapire, Y. Singer, "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers", *Journal of Machine Learning Research 1*, 2000, p. 113–141.

5.1.1 Bornes sur les performances en généralisation des systèmes discriminants multi-classes

Participants: Frédéric Bertrand, Myriam Bertrand, Yannick Darcy, Yann Guermeur, Emmanuel Monfrini.

Un travail qui nous tient particulièrement à cœur en théorie des bornes est l'étude des dimensions de Vapnik-Chervonenkis (VC) étendues pour les systèmes discriminants multi-classes à grande marge. Actuellement, on dispose avec la dimension VC [VC71] d'une mesure de capacité dédiée aux familles de fonctions à valeurs binaires. Les Ψ-dimensions [BDCBHL95] caractérisent la complexité des modèles prenant leurs valeurs dans un ensemble fini tandis que des dimensions comme la dimension fat-shattering [KS90,KS94] permettent d'étudier les systèmes calculant des dichotomies avec une grande marge [ABDCBH97]. Cependant, le cas le plus important en pratique, celui des modèles multi-classes à marge, n'a à notre connaissance été traité que dans [19]. Ce rapport, tout en effectuant quelques avancées, a avant tout souligné la difficulté de la tâche, sur laquelle nous entendons concentrer rapidement une grande partie de nos forces.

Différentes publications ont proposé d'étudier les capacités de généralisation des classifieurs à marge, des machines à noyau et singulièrement des SVM en simplifiant le schéma standard : expression d'un risque garanti faisant intervenir la fonction de croissance ou un nombre de couverture, passage à une dimension VC généralisée par le biais d'un lemme de Sauer-Shelah [VC71,Sau72,She72] étendu et calcul d'une borne sur cette dimension. Parmi celles-ci, on peut en particulier citer [SS04]. Nous travaillons depuis longtemps à l'extension des résultats de Williamson et ses co-auteurs [WSS00,GBSTW02] sur les nombres de couverture des SVM bi-classes (voir en particulier [5, 12]). La seule difficulté qu'il nous reste à surmonter est la détermination

[VC71] V. Vapnik, A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities.", *Theory of Probability and its Applications XVI*, 2, 1971, p. 264–280.

[BDCBHL95] S. Ben-David, N. Cesa-Bianchi, D. Haussler, P. Long, "Characterizations of Learnability for Classes of $\{0,\ldots,n\}$-Valued Functions.", *Journal of Computer and System Sciences 50*, 1995, p. 74–86.

[KS90] M. Kearns, R. Schapire, "Efficient distribution-free learning of probabilistic concepts", *in : Proceedings of the 31st Annual Symposium on Foundations of Computer Science, 1*, IEEE Computer Society Press, p. 382–391, 1990.

[KS94] M. Kearns, R. Schapire, "Efficient Distribution-free Learning of Probabilistic Concepts", *Journal of Computer and System Sciences 48*, 3, 1994, p. 464–497.

[ABDCBH97] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, "Scale-sensitive Dimensions, Uniform Convergence, and Learnability.", *J. ACM 44*, 4, 1997, p. 615–631.

[Sau72] N. Sauer, "On the density of families of sets", *Journal of Combinatorial Theory (A) 13*, 1972, p. 145–147.

[She72] S. Shelah, "A combinatorial problem: Stability and order for models and theories in infinitary languages", *Pacific Journal of Mathematics 41*, 1972, p. 247–261.

[SS04] C. Scovel, I. Steinwart, "Fast Rates for Support Vector Machines", *research report number LA-UR 03-9117*, Los Alamos National Laboratory, 2004.

[WSS00] R. Williamson, A. Smola, B. Schölkopf, "Entropy Numbers of Linear Function Classes", *in : COLT'00*, p. 309–319, 2000.

[GBSTW02] Y. Guo, P. Bartlett, J. Shawe-Taylor, R. Williamson, "Covering Numbers for Support Vector Machines", *IEEE Trans. on Information Theory 48*, 1, 2002, p. 239–250.

de la valeur de la constante universelle du théorème de Maurey-Carl [Car85,CS90] dans le cas dual. Pour mener à bien cette étude (dont l'état d'avancement peut être constaté dans [6]), nous nous appuierons sur la littérature portant sur la théorie des opérateurs [JS82,Sch84,CP88] et les variables aléatoires à valeurs dans des espaces de Banach [LT91].

La thèse d'Olivier Bousquet [Bou02] a connu un fort retentissement dans la communauté apprentissage en mettant en lumière les bénéfices que celle-ci peut tirer des récents développements dans le domaine des inégalités de concentration, de la théorie des processus empiriques et plus particulièrement l'utilisation des moyennes de Rademacher. Nous souhaitons étendre au cas multi-classe la borne dédiée aux SVM faisant intervenir comme mesure empirique de la capacité une fonction du spectre de la matrice de Gram. De manière générale, le travail de synthèse sur la théorie de la discrimination exposé dans [BBL05] sera pour nous une source d'inspiration dans l'optique d'une extension au cas multi-classe des travaux de pointe sur les bornes, standard ou relevant de l'apprentissage PAC-bayésien [McA98].

Plus haut dans cette section, nous avons évoqué nos travaux sur la majoration des nombres de couverture des SVM multi-classes [20] (en passant par les nombres d'entropie de l'opérateur d'évaluation correspondant). L'un des objectifs de ces travaux est la mise au point d'une méthode de sélection de modèle [HTF01,Mas03,BBL05], plus précisément de choix des hyper-paramètres (constante de "marge molle" et paramètres du noyau). Nous souhaitons en particulier nous affranchir des problèmes liés à l'utilisation de la validation croisée, problèmes relevant du dilemme bien connu biais-variance [BG04]. Toujours en sélection de modèle, nous continuerons l'extension des bornes "leave-one-out" [CVBM02]. Dans [3], nous avons déjà proposé une extension de la borne "rayon-marge" qui doit être comparée à celles décrites dans [WXC05].

[Car85] B. Carl, "Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces", *Ann. Inst. Fourier 35*, 3, 1985, p. 79–118.

[CS90] B. Carl, I. Stephani, *Entropy, compactness, and the approximation of operators*, Cambridge University Press, Cambridge, UK, 1990.

[JS82] W. Johnson, G. Schechtman, "Embedding ℓ_p^m into ℓ_1^n", *Acta Math. 149*, 1982, p. 71–85.

[Sch84] C. Schütt, "Entropy Numbers of Diagonal Operators between Symmetric Banach Spaces", *Journal of Approximation Theory 40*, 1984, p. 121–128.

[CP88] B. Carl, A. Pajor, "Gelfand numbers of operators with values in a Hilbert space", *Invent. math. 94*, 1988, p. 479–504.

[LT91] M. Ledoux, M. Talagrand, *Probability in Banach Spaces*, Springer-Verlag, Berlin, 1991.

[Bou02] O. Bousquet, *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*, PdD Thesis, Ecole Polytechnique, 2002.

[BBL05] S. Boucheron, O. Bousquet, G. Lugosi, "Theory of Classification: A Survey of Some Recent Advances", *ESAIM: Probability and Statistics 9*, 2005, p. 323–375.

[McA98] D. McAllester, "Some PAC-Bayesian theorems", *in : COLT'98*, p. 230–234, 1998.

[HTF01] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, *Springer Series in Statistics*, Springer, 2001.

[Mas03] P. Massart, "Concentrations inequalities and model selection", *in : Ecole d'Eté de Probabilités de Saint-Flour XXXIII*, *LNM*, Springer-Verlag, 2003.

[BG04] Y. Bengio, Y. Grandvalet, "No Unbiased Estimator of the Variance of K-Fold Cross-Validation", *Journal of Machine Learning Research 5*, 2004, p. 1089–1105.

[CVBM02] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, "Choosing Multiple Parameters for Support Vector Machines", *Machine Learning 46*, 1, 2002, p. 131–159.

[WXC05] L. Wang, P. Xue, K. Chan, "Generalized Radius-Margin Bounds for Model Selection in

L'étape suivante est la dérivation d'une "span bound" multi-classe. Elle devrait s'effectuer en collaboration avec Liva Ralaivola.

5.1.2 Théorie et pratique des machines à noyau

Participants: Yannick Darcy, Yann Guermeur, Emmanuel Monfrini, Fabienne Thomarat.

En dehors des SVM, la famille des machines à noyau contient plusieurs systèmes discriminants calculant des dichotomies, parmi lesquels certains se distinguent par leurs propriétés statistiques ou leurs performances expérimentales, comme les machines à point bayésien [HGC01] et la "Kernel Projection Machine" (KPM) [BMVZ05]. Nous souhaitons étendre ces machines (borne + algorithme d'apprentissage) au cas multi-classe. L'extension de la KPM devrait en particulier faire l'objet d'une collaboration avec Laurent Zwald.

Ces travaux généralistes (indépendants de toute application particulière) seront complétés par la conception et la mise en œuvre de noyaux dédiés aux séquences. Pour ce faire, nous nous appuierons de manière privilégiée sur les références suivantes : [LEN02,LK03,CHM03].

5.1.3 Apprentissage non supervisé

Participants: Emmanuel Monfrini, Fabienne Thomarat.

Nos contributions en apprentissage non supervisé seront conduites dans l'optique de concevoir ou améliorer des méthodes phylogénétiques [Fel04] (voir à ce sujet la section 6). Elles se développeront prioritairement suivant deux axes. D'une part, nous étudierons la robustesse des méthodes de classification aux données lacunaires et bruitées [Wie03,PSB+04]. D'autre part, nous développerons des méthodes de classification fondées sur l'emploi d'un noyau reproduisant et adaptées aux grandes masses de données [26]. Ces travaux profiteront des recherches d'Emmanuel Monfrini sur les distributions de mélanges pour l'apprentissage non supervisé [22].

Multi-class SVMs", *research report*, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, 2005.

[HGC01] R. Herbrich, T. Graepel, C. Campbell, "Bayes Point Machines", *Journal of Machine Learning Research 1*, 2001, p. 245–279.

[BMVZ05] G. Blanchard, P. Massart, R. Vert, L. Zwald, "Kernel Projection Machine: a New Tool for Pattern Recognition", *in : NIPS'17*, p. 1649–1656, 2005.

[LEN02] C. Leslie, E. Eskin, W. S. Noble, "The spectrum kernel: A string kernel for SVM protein classification", *Proceedings of the Pacific Symposium on Biocomputing*, 2002.

[LK03] C. Leslie, R. Kuang, "Fast Kernels for Inexact String Matching", *in : 16th Annual Conference on Learning Theory (COLT'03)*, 2003.

[CHM03] C. Cortes, P. Haffner, M. Mohri, "Positive Definite Rational Kernels", *in : 16th Annual Conference on Learning Theory (COLT'03)*, p. 41–56, 2003.

[Fel04] J. Felsenstein, *Inferring Phylogenies*, Sinauer, 2004.

[Wie03] J. Wiens, "Missing data, incomplete taxa, and phylogenetic accuracy", *Syst. Biol. 52*, 4, 2003, p. 528–538.

[PSB+04] H. Philippe, E. Snell, E. Bapteste, P. Lopez, P. Holland, D. Casane, "Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments", *Molecular Biology and Evolution 21*, 9, 2004, p. 1740–1752.

5.2 Biologie computationnelle

En biologie computationnelle, nous souhaitons continuer à travailler principalement sur des problèmes de traitement de séquences. Cette thématique de la reconnaissance des formes, encore largement ouverte, est celle où notre expertise se trouve être la plus grande, et pour laquelle nous cherchons de manière privilégiée à développer des solutions à la fois génériques et opérationnelles. La majorité des problèmes auxquels s'intéressent les équipes de biologistes avec lesquelles nos liens, établis de longue date, sont les plus forts, relève de ce domaine.

5.2.1 Ingénierie du noyau

Participants: Yannick Darcy, Yann Guermeur, Emmanuel Monfrini, Fabienne Thomarat.

Les travaux que nous effectuerons dans ce domaine sont très liés à ceux évoqués dans les sous-sections 5.1.2 et 5.1.3. Un axe que nous privilégierons pour le traitement des séquences biologiques est celui des noyaux fondés sur des HMM [JDH99,Wat99,Hau99]. En comparaison des autres types de "string kernels" (voir par exemple [LSST+02]), ceux-ci apparaissent, au moins en principe, mieux adaptés à la prise en compte des phénomènes de l'évolution biologique que sont les substitutions ainsi que les insertions/délétions. Leur utilisation soulève cependant des problèmes techniques importants (faiblesse du lien existant entre qualité de la modélisation et pouvoir discriminant, mauvais conditionnement de la matrice de Gram...), que nous nous attacherons à résoudre.

5.2.2 Développement d'architectures hybrides, intégrant systèmes discriminants et génératifs

Participants: Yann Guermeur.

Ces travaux, en lien direct avec ceux évoqués dans la sous-section précédente, nous permettront en particulier de poursuivre le développement de la méthode de prédiction de la structure secondaire des protéines globulaires initialement décrite dans [16]. Il est à noter que l'idée d'effectuer un post-traitement des sorties de SVM au moyen de HMM, idée que nous avons introduite à la fin des années 90 en prédiction de la stucture secondaire [17], a récemment prouvé son utilité dans d'autres domaines de la reconnaissance des formes, en particulier en parole [SR04].

[JDH99] T. Jaakola, M. Diekhans, D. Haussler, "Using the Fisher kernel method to detect remote protein homologies", *in : ISMB'99*, p. 149–158, 1999.

[Wat99] C. Watkins, "Dynamic Alignment Kernels", *research report number CSD-TR-98-11*, Department of Computer Science, Royal Holloway, University of London, 1999.

[Hau99] D. Haussler, "Convolution Kernels on Discrete Structures", *research report number UCSD-CRL-99-10*, Departement of Computer Science, University of California at Santa Cruz, 1999.

[LSST+02] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins, "Text Classification using String Kernels", *Journal of Machine learning Research 2*, 2002, p. 419–444.

[SR04] J. Stadermann, G. Rigoli, "A Hybrid SVM/HMM Acoustic Modeling Approach for Automatic Speech Recognition", *in : INTERSPEECH 2004 - ICSLP*, p. 661–664, 2004.

6 Domaines d'application

Nos travaux peuvent trouver des applications dans pratiquement tous les domaines de la reconnaissance des formes : traitement automatique de la parole, des corpus de textes, des images, de l'écriture cursive... Ceci nous ouvre des perspectives de collaborations avec de nombreuses équipes du LORIA. Cependant, notre domaine d'application privilégié, voire exclusif dans un premier temps, sera celui du traitement de séquences biologiques. Le choix des problèmes sur lesquels travaillera un membre de l'équipe sera motivé de manière différente suivant son profil. Tandis que les biologistes mettront naturellement en œuvre dans ce cadre leur propre projet de recherche, les informaticiens et statisticiens se fonderont sur les collaborations qu'ils auront avec des biologistes. Dans les deux prochaines années, compte tenu des collaborations en cours (voir la section 8.2), trois grandes tâches devraient occuper la majeure partie de nos forces :

- l'assignation et la prédiction des structures des protéines,
- l'étude du phénomène d'épissage alternatif,
- la phylogénie moléculaire.

7 Logiciels

L'ensemble de nos travaux, que ce soit en théorie de l'apprentissage ou en bioinformatique, nous conduira de manière naturelle à développer des logiciels qui seront largement diffusés auprès des deux communautés.

7.1 Apprentissage automatique

En apprentissage, nous continuerons à développer notre logiciel de SVM multi-classes [14]. Le développement se fera essentiellement par ajout de fonctionnalités nouvelles, comme l'intégration de la méthode de sélection de modèle proposée par Yannick Darcy dans le cadre de ses travaux de thèse [3]. Nous espérons reprendre une ancienne collaboration avec Anatoli Juditsky, responsable de l'équipe "Statistique et modélisation stochastique" (SMS) au LMC, à Grenoble, collaboration portant sur la recherche d'une programmation efficace de l'algorithme d'apprentissage des M-SVM. Nous pourrions aussi profiter dans ce domaine de l'aide apportée par des membres du projet GRAAL, dans le cadre du programme Décrypthon évoqué plus bas. Nous souhaitons également que l'étude de la machine à point bayésien et de la KPM multi-classes donne lieu à la réalisation de programmes. Toutes les applications relevant de cette section, c'est-à-dire celles n'étant pas dédiées à une application bioinformatique, seront en particulier diffusées à partir du site des "Kernel Machines" `http://www.kernel-machines.org/`.

7.2 Bioinformatique

Notre longue collaboration avec l'équipe de Gilbert Deléage a donné naissance à un ensemble d'applications librement accessibles depuis le serveur d'analyse de séquences protéiques du Pôle Bioinformatique Lyonnais (PBIL), `http://pbil.univ-lyon1.fr/` et plus précisément le site NPS@, `http://npsa-pbil.ibcp.fr`. Le logiciel "AmphipaSeeK" mettant en œuvre la

méthode de prédiction des ancrages membranaires interfaciaux dans les protéines membranaires monotopiques introduite dans [25] vient d'être installé sur ce même serveur, à l'adresse suivante :
`http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_amphipaseek.html`.
Ce serveur est pour nous la destination première de nos réalisations en bioinformatique.

8 Participation à des projets au niveau local, national et international

Du fait de la nature de notre projet scientifique, et du domaine d'application privilégié des résultats produits, notre ambition sera de renforcer nos positions au sein de deux communautés, celle de l'apprentissage automatique et celle de la bioinformatique, ceci à tous les niveaux où elles se structurent.

8.1 Apprentissage automatique

Le domaine de l'apprentissage est celui où nos positions sont actuellement les mieux établies. Nous avons déjà par le passé assumé des responsabilités d'animation scientifique sur ce sujet, comme la coordination de groupes de travail d'actions spécifiques du CNRS et d'un projet d'une ACI.

Nous avons également été membres de réseaux d'excellence européens, comme NeuroCOLT : `http://www.neurocolt.com/`.

Dans un avenir proche, nous souhaitons inscrire nos travaux en apprentissage dans trois contextes :

- des collaborations avec certaines des équipes du LORIA dont l'activité possède une composante apprentissage : Magrit, Maia, Orpailleur, Parole, Qgar ou READ,
- la "Fédération des Equipes de Recherche en Apprentissage" (FERA) `http://www.lri.fr/~proml/wiki/index.php/Main_Page`,
- le réseau d'excellence européen "Pattern Analysis, Statistical Modelling and Computational Learning" (PASCAL) `http://www.pascal-network.org/`.

Le renforcement des liens de l'équipe avec PASCAL est pour nous un objectif prioritaire, même si sa réalisation doit prendre du temps. En effet, ce réseau constitue la référence au niveau européen en matière de théorie statistique de l'apprentissage, et y être associé facilite grandement les collaborations avec les équipes, européennes ou non, dont les thèmes de recherche sont proches des nôtres. Le rapprochement est déjà en cours, au moins à titre individuel, au travers par exemple d'une participation à l'organisation du challenge théorique "Type I and type II errors for multiple simultaneous hypothesis" qui a été lancé au début de l'année `http://www.lri.fr/~teytaud/risq/risq.html`. Ce challenge renforce également nos liens avec le projet "Thème Apprentissage et Optimisation" (TAO) de l'INRIA Futurs puisqu'il est porté par Olivier Teytaud.

8.2 Bioinformatique

En bioinformatique, dans un premier temps, notre implication se situera plutôt au niveau régional et national. La raison en est le nombre de collaborations de ce type déjà en cours, que nous devons mener à terme avant d'envisager de prendre des engagements nouveaux. Ces collaborations sont les suivantes :

1. Projet "Apprentissage automatique appliqué à la prédiction de la structure tertiaire des protéines" (GENOTO3D) financé par l'ACI "Masses de Données" (Yann Guermeur coordinateur) `http://www.loria.fr/~guermeur/ACIMD/`.
2. Projet "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, est financé pour 18 mois par le programme Décrypthon `http://www.decrypthon.fr/`. Il fait également l'objet d'une opération du thème "Bioinformatique et applications à la génomique" du "PRST Intelligence Logicielle".
3. Projet "Modélisation de la protéine FAK (Focal Adhesion Kinase) en vue de l'identification de molécules anti-métastases", en collaboration avec l'"équipe de Dynamique des Assemblages Membranaires" (eDAM) du laboratoire "Structure et Réactivité des Systèmes Moléculaires Complexes", UMR 7565 à Nancy. Ce projet est une opération du thème "Bioinformatique et applications à la génomique" du "PRST Intelligence Logicielle". Il est également financé par l'ANR, dans le cadre du projet "Tyrosines kinases de la famille FAK. Bases structurales de la régulation et de la localisation intracellulaire", retenu par l'ANR non-thématique pour les années 2006 à 2008. Il devrait enfin servir de base pour la mise en place d'une collaboration internationale dont l'équipe actuelle d'Alexander Bockmayr à Berlin serait l'un des partenaires.

A cette liste viennent s'ajouter des participations à différents groupes de travail, comme le "Groupe de travail pluridisciplinaire sur la structure et la fonction des ARN" (AReNa) soutenu par l'ACI IMPBio. Le co-encadrement avec Nadir Mrabet, Professeur de Biologie à la Faculté de Médecine de l'Université Nancy 1, du stage de Master de Levoly Fani a également posé les bases d'une nouvelle collaboration en biologie structurale prédictive. Il s'agira de poursuivre le développement du logiciel Spiralix, destiné à l'identification des structures secondaires régulières et à la comparaison des structures tertiaires associées.

9 Autres collaborations

En plus des collaborations relevant d'un cadre précis (associées à un financement), évoquées plus haut, nous souhaitons poursuivre ou relancer nos collaborations informelles avec diverses équipes en apprentissage, statistique, bioinformatique et biostatistique. Parmi celles-ci, on peut en particulier citer l'équipe connexionniste du LIP6, le projet TAO déjà évoqué dans la section 8.1, le Département "Inférence empirique pour l'apprentissage machine et la perception", dirigé par Bernhard Schölkopf, au "MPI for Biological Cybernetics", à Tübingen, ainsi que le "Laboratoire de Bioinformatique et RMN Structurales" (LBRS) de l'IBCP, à Lyon, ou le groupe de Pierre Baldi à l'Université de Californie à Irvine (UCI). Nous souhaitons initier une

collaboration avec l'équipe de Gérard Biau, à l'"Institut de Mathématiques et de Modélisation de Montpellier" (I3M), en apprentissage des données fonctionnelles [BBW05]. Ce thème, en plein essort, présente pour nous au moins deux attraits : d'une part, il constitue un domaine de choix pour le développement de méthodes à noyau, d'autre part, il trouve de nombreuses applications en traitement des séquences biologiques. Toujours en collaboration avec des équipes de Montpellier, cette fois l'équipe d'Olivier Gascuel, au LIRMM, et celle d'Emmanuel Douzery, à l'ISEM, nous poursuivrons nos travaux sur l'impact des données manquantes en phylogénie moléculaire.

Remerciements
Les membres de l'équipe souhaitent exprimer leur profonde gratitude à Madame Michèle Sebag, Monsieur le Professeur Gérard Biau, Monsieur le Professeur Jean-Paul Haton et Monsieur Olivier Teytaud pour leur aide dans la conception et la mise en forme de ce projet de recherche.

10 Bibliographie

Principales publications de l'équipe au cours des dernières années

[1] M. Bertrand, B. Boulanger, A. Gilbert, B. Govaerts, D. Walthere, "Risk management for analytical methods based on the total error concept : conciliating the objectives of the pre-study and in-study validation phases", *Chemometrics and Intelligent Laboratory Systems*, 2006, (à paraître).

[2] A. Brehm, D. Harris, C. Alves, J. Jesus, F. Thomarat, L. Vicente, "Structure and evolution of the mitochondrial DNA complete control region in the lizard *Lacerta dugesii* (Lacertidae, Sauria)", *J Mol Evol 56*, 2003, p. 46–53.

[3] Y. Darcy, E. Monfrini, Y. Guermeur, "Borne "rayon-marge" sur l'erreur "leave-one-out" des SVM multi-classes", *in : JdS'06*, 2006.

[4] J.-C. Deville, M. .Maumy, "La méthodologie de Morgoat : enquête tourisme en Bretagne", *in : Méthodes d'Enquêtes et Sondages, Pratiques européenne et nord-américaine*, P. Lavallée and L.-P. Rivest (editors), Dunod, 2006, p. 393–398.

[5] A. Elisseeff, Y. Guermeur, H. Paugam-Moisy, "Margin error and generalization capabilities of multi-class discriminant models", *research report number NC-TR-99-051-R*, NeuroCOLT2, 1999, (revised in 2001).

[6] Y. Guermeur, M. Bertrand, F. Sur, "Notes sur le "théorème de Maurey-Carl"", *research report*, INRIA, 2006, (en préparation).

[7] Y. Guermeur, A. Elisseeff, H. Paugam-Moisy, "Estimating the sample complexity of a multi-class discriminant model", *in : ICANN'99*, IEE, p. 310–315, 1999.

[8] Y. Guermeur, A. Elisseeff, H. Paugam-Moisy, "SVM multiclasses", *in : Support Vector Machines et autres méthodes à noyau*, S. Canu, C. Richard, M. Davy, and A. Rakotomamonjy (editors), Hermès, 2006, (à paraître).

[9] Y. Guermeur, A. Elisseeff, D. Zelus, "A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers", *Applied Stochastic Models in Business and Industry 21*, 2, 2005, p. 199–214.

[BBW05] G. Biau, F. Bunea, M. Wegkamp, "Functional Classification in Hilbert Spaces", *IEEE Transactions on Information Theory 51*, 2005, p. 2163–2172.

[10] Y. Guermeur, C. Geourjon, P. Gallinari, G. Deléage, "Improved performance in protein secondary structure prediction by inhomogeneous score combination", *Bioinformatics 15*, 5, 1999, p. 413–421.

[11] Y. Guermeur, A. Lifchitz, R. Vert, "A kernel for protein secondary structure prediction", *in : Kernel Methods in Computational Biology*, B. Schölkopf, K. Tsuda, and J.-P. Vert (editors), The MIT Press, 2004, p. 193–206.

[12] Y. Guermeur, M. Maumy, F. Sur, "Model selection for multi-class SVMs", *in : ASMDA'05*, p. 507–516, 2005.

[13] Y. Guermeur, H. Paugam-Moisy, "Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines", *in : Apprentissage Automatique*, M. Sebban and G. Venturini (editors), Hermès, 1999, p. 109–138.

[14] Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy, P. Baldi, "Combining Protein Secondary Structure Prediction Models with Ensemble Methods of Optimal Complexity", *Neurocomputing 56C*, 2004, p. 305–327.

[15] Y. Guermeur, O. Teytaud, "Estimation et contrôle des performances en généralisation des réseaux de neurones", *in : Apprentissage Connexionniste*, Y. Bennani (editor), Hermès, 2006, ch. 10, p. 279–342.

[16] Y. Guermeur, *Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines*, PhD Thesis, Université Paris 6, 1997.

[17] Y. Guermeur, "Combining discriminant models with new multi-class SVMs", *research report number NC2-TR-2000-086*, NeuroCOLT2, 2000.

[18] Y. Guermeur, "Combining discriminant models with new multi-class SVMs", *Pattern Analysis and Applications 5*, 2, 2002, p. 168–179.

[19] Y. Guermeur, "Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions", *Research Report RR-5314*, INRIA, September 2004, `http://www.inria.fr/rrrt/rr-5314.html`.

[20] Y. Guermeur, "SVM multiclasses", *in : Support Vector Machines et autres méthodes à noyau*, S. Canu, C. Richard, M. Davy, and A. Rakotomamonjy (editors), Hermès, 2007, (à paraître).

[21] M. Katinka, S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. E. Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, C. Vivares, "Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*", *Nature 414*, 2001, p. 450–453.

[22] E. Monfrini, W. Pieczynski, "Estimation de mélanges généralisés dans les arbres de Markov cachés, application à la segmentation de cartons d'orgues de barbarie", *Traitement du Signal 22*, 2, 2005, p. 135–147.

[23] E. Monfrini, "Unicité de la méthode des moments pour le mélange de deux distributions normales", *C. R. Acad. Sci. Paris, Ser. I 336*, 2003, p. 89–94.

[24] E. Monfrini, "Une méthode des moments stable pour les mélanges de deux distributions normales", *Rev. Roumaine Math. Pures Appl. 49*, 1, 2004, p. 45–62.

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

[26] F. Thomarat, C. Vivares, M. Gouy, "Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes", *J Mol Evol. 59*, 6, 2004, p. 780–791.

[27] C. Vivares, M. Gouy, F. Thomarat, G. Metenier, "Functional and evolutionary analysis of a eukaryotic parasitic genome", *Curr Opin Microbiol 5*, 2002, p. 499–505.