Etude comparée des performances de SVM multi-classes en prédiction de la structure secondaire des protéines

Y. Guermeur

LORIA-CNRS, équipe ABC
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy cedex
Yann.Guermeur@loria.fr
http://www.loria.fr/~guermeur

Résumé. Les SVM bi-classes, introduites en bioinformatique à la fin des années 90, font aujourd'hui référence pour de nombreux problèmes de traitement de séquences biologiques. Les SVM multi-classes, de conception plus récente, sont progressivement appliquées à ces problèmes, singulièrement en biologie structurale prédictive. Dans cet article, nous proposons une étude comparée des performances de trois SVM multi-classes en prédiction de la structure secondaire des protéines. Les modèles impliqués sont celui de Weston et Watkins, celui de Lee et co-auteurs ainsi qu'une nouvelle machine nommée M-SVM^2. Cette étude se conçoit comme une étape dans la mise au point d'une méthode de prédiction hybride, intégrant systèmes discriminants et génératifs et s'appuyant sur une approche hiérarchique du problème.

1 Introduction

La biologie moléculaire est un domaine d'application de choix pour les modèles de l'apprentissage artificiel. Si les problèmes de discrimination qu'elle propose font intervenir des données de natures très diverses, un grand nombre d'entre eux, souvent parmi les plus importants, relèvent du traitement de séquences. C'est en particulier le cas des problèmes de biologie structurale prédictive. L'un des plus anciens, qui a résisté à quarante années de recherches intensives, est la prédiction *ab initio* du repliement des protéines globulaires. Il est ordinairement abordé par le biais d'une approche du type diviser pour régner faisant intervenir une sous-tâche nommée prédiction de la structure secondaire. Du point de vue de l'apprentissage artificiel, cette tâche de discrimination à catégories multiples est d'un intérêt majeur, ceci à plus d'un titre. Pour nous restreindre à ce qui fera l'objet de cet article, il convient d'illustrer cet intérêt en précisant que les principaux modèles connexionnistes discriminants ont été appliqués en prédiction de la structure secondaire, de même que les machines à vecteurs support (SVM). Dans ce domaine, les perceptrons multi-couches (PMC) et leurs variantes récurrentes constituent l'état de l'art depuis environ vingt ans. Ils ont été rejoints récemment par les SVM. Le choix de ce problème est donc particulièrement indiqué pour réaliser une étude comparative de SVM multi-classes (M-SVM). C'est précisément ce que propose cet article. Les trois machines considérées sont la M-SVM de Weston et Watkins (1998), celle de Lee et al. (2004)

et une nouvelle machine nommée M-SVM^2 (Monfrini et Guermeur, 2008). L'objectif final est la mise au point d'une méthode de prédiction hybride, intégrant systèmes discriminants et génératifs et s'appuyant sur une approche hiérarchique du problème.

Le plan de l'article est le suivant. La section 2 présente le problème de la prédiction de la structure secondaire des protéines, ainsi que les méthodes constituant l'état de l'art. La description des trois M-SVM impliquées dans l'étude comparative fait l'objet de la section 3. Le noyau à fonction radiale de base (RBF) qu'elles utilisent toutes est présenté dans la section 4. La question de la sélection de modèle est étudiée dans la section 5. La section 6 est dédiée à l'étude comparative proprement dite. Enfin, la section 7 expose nos conclusions et nos perspectives de recherche.

2 Prédiction de la structure secondaire des protéines

Dans cette section, nous présentons le problème de traitement de séquences biologiques sur lequel porte cette étude comparative.

2.1 Présentation du problème

Connaître la structure d'une protéine est un prérequis pour comprendre précisément sa fonction. Les projets de séquençage à grande échelle qui se sont multipliés ces dernières années ont permis d'obtenir les séquences d'un très grand nombre de gènes et par suite celles d'un très grand nombre de protéines. Le phénomène a été accéléré par l'apparition de nouvelles techniques de séquençage à la fois rapides et à bas coût. Malheureusement, le nombre de structures connues n'a pas suivi la même progression. En effet, les méthodes expérimentales disponibles pour déterminer la structure tridimensionnelle (*tertiaire*), la cristallographie par rayons X ou radiocristallographie et la spectroscopie par résonance magnétique nucléaire (RMN) demandent beaucoup d'efforts ou sont tout simplement inutilisables. Ainsi, certaines protéines ne cristallisant pas ne peuvent relever de la radiocristallographie. De ce fait, prédire la structure tertiaire des protéines *ab initio*, c'est-à-dire à partir de la seule séquence (*structure primaire*), est devenu l'un des problèmes centraux de la biologie structurale. Au début des années 60, Anfinsen proposa son "hypothèse thermodynamique" (Epstein et al., 1963), impliquant que la séquence protéique contient suffisamment d'information pour garantir un repliement correct à partir d'un vaste ensemble d'états dépliés. Cette hypothèse s'appuyait en particulier sur des expériences de "dénaturation-renaturation" (Anfinsen et al., 1961). Si le problème considéré peut donc théoriquement être résolu, les difficultés pratiques, mises en évidence par exemple par Karplus et Petsko (1990), sont telles qu'il est rarement abordé de manière directe, mais plutôt au travers d'une approche du type diviser pour régner. Dans ce contexte, une étape intermédiaire utile est la prédiction de la *structure secondaire*, qui représente un moyen de simplifier le problème en projetant la structure 3D très complexe sur une dimension, c'est-à-dire sur une succession d'états conformationnels associés à chaque résidu (acide aminé) de la séquence. La structure secondaire d'une protéine est constituée par les motifs réguliers (périodiques) et répétés du repliement de son épine dorsale. Les deux éléments structuraux de base sont l'*hélice alpha* et le *brin bêta*, auxquels s'ajoute une transition apériodique, le *coil*. La figure 1 propose une représentation schématique de la structure secondaire de

la protéine G (Derrick et Wigley, 1994), représentation obtenue avec le logiciel RasMol (Sayle et Milner-White, 1995).

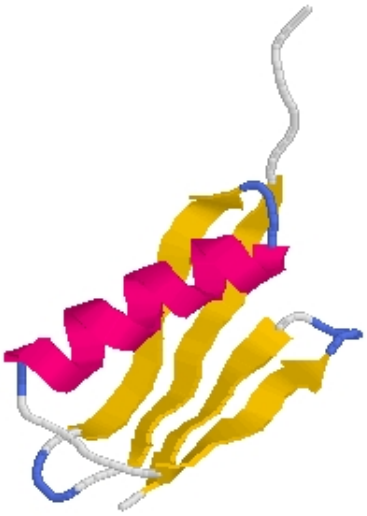

FIG. 1 – *Représentation schématique des éléments structuraux de la protéine G. Cette structure est composée de deux parties principales : une hélice alpha, matérialisée en rouge, et un* feuillet bêta *constitué de quatre brins, en jaune.*

Si la prédiction de la structure secondaire peut être utilisée pour fournir des contraintes aux méthodes de prédiction *ab initio* de la structure tertiaire, elle peut également jouer un rôle dans la mise en œuvre des deux autres grandes stratégies de prédiction de la structure tertiaire que sont la reconnaissance de repliement (*threading*) (voir par exemple Russell et al., 1996; Rost et al., 1997) et la *modélisation par homologie/analogie* (voir par exemple Boscott et al., 1993; Combet et al., 2002). Considérée comme une tâche de reconnaissance des formes, elle se présente sous la forme d'un problème de discrimination à trois catégories consistant à affecter à chaque résidu de la séquence son état conformationnel, en hélice α, brin β ou structure apériodique (coil).

2.2 Etat de l'art

Les premiers travaux en prédiction de la structure secondaire datent de la fin des années 60. Depuis lors, ce problème a fait l'objet de recherches intensives. Il est possible de classer en trois grandes familles les méthodes qui ont été mises en œuvre pour le traiter. Historiquement, les méthodes les plus anciennes sont celles fondées sur l'exploitation de propriétés physico-chimiques (Lim, 1974b,a) et celles dites "statistiques" (Chou et Fasman, 1978; Gibrat et al., 1987), reposant principalement sur l'estimation des probabilités conditionnelles des états conformationnels à partir de statistiques d'ordres un ou deux calculées sur de petits peptides. Elles ont progressé lentement au cours des années 80-90 (Gaboriaud et al., 1987; Solovyev et

Salamov, 1994; Geourjon et Deléage, 1995), jusqu'au moment où elles ont été pratiquement supplantées par des méthodes issues de l'apprentissage numérique. Les progrès les plus spectaculaires ont résulté de l'introduction dans le domaine de PMC. Dans un premier temps, une transposition relativement simple du système NETtalk a permis à Qian et Sejnowski (1988) de faire passer le taux de reconnaissance (taux de résidus correctement classés) d'environ 60% à plus de 64%. Leur architecture doit être évoquée, car elle a été pratiquement systématiquement reprise dans les travaux ultérieurs. Deux PMC sont utilisés en cascade. Le premier, dit "séquence-structure", effectue la prédiction de la structure à partir de la séquence, le second, dit "structure-structure", lissant la prédiction initiale, en prenant en compte le fait que les états conformationnels des résidus consécutifs sont fortement corrélés. Le remplacement, en entrée du PMC "séquence-structure", de la simple structure primaire par un *profil d'alignement*, a eu pour conséquence un accroissement supplémentaire du taux de reconnaissance de plus de 6%, permettant de dépasser pour la première fois la frontière des 70% de résidus correctement classés. Ainsi, la méthode PHD de Rost et Sander (1993), à travers ses développements successifs (Rost et Sander, 1994), a constitué l'état de l'art pendant la plus grande partie des années 90. Les alignements qu'elle utilise sont issus de la base HSSP (Sander et Schneider, 1991; Dodge et al., 1998). Pour un alignement donné, le profil est obtenu en calculant à chaque position les fréquences d'apparition des vingt acides aminés. Parallèlement, les systèmes mettant en œuvre des modèles de Markov cachés (HMM), introduits par Asai et al. (1993), ont tiré profit, comme les autres, de l'accroissement de la taille de la *protein data bank* (PDB) (Bernstein et al., 1977; Berman et al., 2000) pour prendre en compte de manière de plus en plus fine les règles syntaxiques régissant l'agencement des éléments structuraux. En la matière, la contribution la plus aboutie est probablement constituée par les travaux de Martin et al. (2005); Martin (2005). Actuellement, les meilleurs systèmes de prédiction sont des modèles connexionnistes complexes, intégrant jusqu'à plusieurs centaines de réseaux, à propagation avant ou récurrents (Jones, 1999; Baldi et al., 1999; Cuff et Barton, 2000; Petersen et al., 2000; Pollastri et al., 2002). Si la combinaison de modèles permet d'améliorer significativement les performances (voir aussi Cuff et Barton, 1999), celles-ci dépendent principalement de la qualité des alignements exploités. Depuis son introduction par Jones (1999) dans le cadre de la méthode de prédiction PSIPRED, l'emploi de profils d'alignements issus de l'algorithme PSI-BLAST (Altschul et al., 1997; Altschul et Koonin, 1998) et s'appuyant sur une matrice de scores dépendant de la position (PSSM), est quasi-exclusif. A l'inverse, l'importance de la nature des réseaux neuronaux apparaît secondaire. On n'explique pas le fait que ceux-ci soient souvent largement sur-paramétrés, sans que cela n'entraîne pour autant de phénomène de sur-apprentissage (Riis et Krogh, 1996; Guermeur, 1997). Depuis plusieurs années, les taux de reconnaissance les plus élevés rapportés saturent aux environs des 80% de résidus bien classés. De ce fait, le dernier grand article de synthèse que Rost a écrit sur la prédiction de la structure secondaire, (Rost, 2001), demeure aujourd'hui encore d'actualité.

La communauté de la biologie structurale prédictive est en attente d'un progrès du même ordre que celui ayant résulté de l'emploi de PMC ou d'alignements multiples. Un autre phénomène vient expliquer le ralentissement des progrès effectués. Nous avons vu dans la section 2.1 qu'il existe trois grandes approches pour réaliser la prédiction de la structure tertiaire : la *modélisation par homologie* (Sali, 1995), le *threading* (Lemer et al., 1995; Marin et al., 2002) et l'approche *ab initio* (Jones, 1997; Hardin et al., 2002). Ce sont principalement les méthodes de prédiction *ab initio* qui utilisent la prédiction de la structure secondaire. Or, on ne fait appel

à elles que lorsque la modélisation par homologie et le *threading* ont échoué. La conjugaison de deux facteurs positifs : les progrès méthodologiques et l'accroissement continuel des tailles des bases de référence, diminue sensiblement la fréquence de cette situation. De plus, la méthode de prédiction *ab initio* actuellement la plus performante, ROSETTA (Simons et al., 1999), n'est pas fondée sur la prédiction de la structure secondaire. Au contraire, elle a été utilisée par Meiler et Baker (2003) afin d'améliorer cette prédiction. La tâche qui nous intéresse a donc perdu, au cours des années, un peu de son importance. Elle pourrait revenir au devant de la scène si une avancée significative avait de nouveau lieu. Naturellement, il est plaisant d'imaginer qu'une telle avancée puisse résulter de l'emploi de méthodes à noyau, et plus particulièrement de M-SVM.

A notre connaissance, la première application d'une SVM, en l'occurence multi-classe, en prédiction de la structure secondaire, était une combinaison de modèles (Guermeur, 2000) (voir aussi Guermeur, 2002). Ces travaux, qui ont produit un gain de performance statistiquement significatif par rapport aux méthodes d'ensemble habituellement utilisées dans ce contexte, ont été poursuivis dans (Guermeur et al., 2004), avec le même succès. Parallèlement, Hua et Sun (2001) ont été les premiers à aborder directement le problème. Pour ce faire, ils ont appliqué à des profils d'alignements multiples des SVM bi-classes à noyau gaussien sphérique, combinées au moyen de la méthode de décomposition "un contre tous" (Rifkin et Klautau, 2004) ainsi que d'autres méthodes élémentaires fondées sur un graphe de décision. Comme dans le cas de la méthode PHD, leurs alignements multiples étaient issus de la base HSSP, les profils résultant du calcul en chaque position des fréquences des différents acides aminés. Leur conclusion est que leur approche permet d'obtenir une très bonne valeur du coefficient Sov (Rost et al., 1994; Zemla et al., 1999) global : 76.2%. La première étude décrivant l'application de SVM bi-classes à des profils d'alignements issus de PSI-BLAST (toujours dans le cadre de la mise en œuvre de méthodes de décomposition) est due à Kim et Park (2003). Des travaux similaires ont été effectués par Wang et al. (2004). La principale différence par rapport aux précédents réside dans l'utilisation d'un codage physico-chimique des acides aminés. Le noyau, à l'inverse, demeure le même. Le taux de reconnaissance annoncé, 78.4%, est comparable à celui des meilleurs systèmes connexionnistes. Suivant l'usage commun, adopté par exemple par le challenge *critical assessment of techniques for protein structure prediction* (CASP), nous avons considéré jusqu'à présent que la prédiction de la structure secondaire était un problème de discrimination à trois catégories. En fait, les programmes d'assignation déterminant la structure secondaire à partir de la structure 3D considèrent plus d'états conformationnels. Ainsi, le programme DSSP (Kabsch et Sander, 1983), le plus utilisé en pratique, en considère huit, parmi lesquels celui correspondant aux coudes β, désigné par le symbole T, qu'une classification en seulement trois catégories associe à la structure apériodique (voir par exemple Cuff et Barton, 1999). Ces coudes β sont prédits au moyen d'une SVM par Zhang et al. (2005). Les prédicteurs utilisés sont à nouveau ceux correspondant au profil d'alignement obtenu par application de PSI-BLAST plus la structure secondaire prédite par PSIPRED. Une fois de plus, les performances sont encourageantes, alors même que le noyau utilisé repose sur une simple gaussienne sphérique. On trouve plusieurs autres applications des SVM en prédiction de la structure secondaire utilisant des SVM bi-classes et le noyau gaussien standard (voir par exemple Ward et al., 2003; Guo et al., 2004). A l'inverse, à notre connaissance, une seule autre équipe a appliqué une M-SVM à ce problème. Nguyen et Rajapakse (2003, 2005) reprennent l'architecture introduite par Qian et Sejnowski, consistant à mettre en œuvre en cascade deux

systèmes discriminants, l'un "séquence-structure" et l'autre "structure-structure". Les PMC de la méthode originale sont remplacés par des M-SVM de Crammer et Singer (2001). Ici encore, le noyau utilisé est le noyau gaussien standard. La section suivante présente les M-SVM que nous avons mises en œuvre.

3 Trois modèles de SVM multi-classes

Une introduction générale aux M-SVM est un préalable à la caractérisation des trois machines d'intérêt dans cet article.

3.1 Présentation générale

Il convient en premier lieu de définir les familles de fonctions sur lesquelles s'appuient toutes les M-SVM.

3.1.1 Familles de fonctions $\mathcal{H}$

Soit $\mathcal{X}$ un espace quelconque et κ un noyau continu, symétrique, semi-défini positif ou *noyau de Mercer (1909)* sur $\mathcal{X}$. Soit $\left(H_\kappa, \langle\cdot,\cdot\rangle_{H_\kappa}\right)$ l'espace de Hilbert à noyau reproduisant (RKHS) (Berlinet et Thomas-Agnan, 2004) correspondant. L'existence et l'unicité de $\left(H_\kappa, \langle\cdot,\cdot\rangle_{H_\kappa}\right)$ sont assurées par le théorème de Moore-Aronszajn (Aronszajn, 1950) (voir aussi Berlinet et Thomas-Agnan, 2004, théorème 3). Le théorème de Mercer nous apprend qu'il existe une application Φ de $\mathcal{X}$ dans un espace de Hilbert $\left(E_{\Phi(\mathcal{X})}, \langle\cdot,\cdot\rangle\right)$ satisfaisant :

$$\forall\,(x,x')\in\mathcal{X}^2,\ \kappa\,(x,x') = \langle\Phi\,(x)\,,\Phi\,(x')\rangle. \tag{1}$$

Par abus de langage, on parlera de l'*espace de représentation* (*feature space*) pour évoquer l'un quelconque des espaces de Hilbert $\left(E_{\Phi(\mathcal{X})}, \langle\cdot,\cdot\rangle\right)$ ainsi définis. Q étant un entier supérieur ou égal à deux, du fait même de la définition d'un RKHS, $\mathcal{H} = \left(\left(H_\kappa, \langle\cdot,\cdot\rangle_{H_\kappa}\right) + \{1\}\right)^Q$ est la famille des fonctions à valeurs vectorielles $h = (h_k)_{1\leq k\leq Q}$ dont les fonctions composantes sont les combinaisons affines de tailles finies m_k de la forme :

$$h_k(\cdot) = \sum_{i=1}^{m_k} \beta_{ik}\kappa\,(x_{ik},\cdot) + b_k,$$

où les x_{ik} sont des éléments de $\mathcal{X}$ (les β_{ik} et b_k sont des scalaires), ainsi que les limites de ces fonctions lorsque les ensembles $\{x_{ik}\ : 1\leq i\leq m_k\}$ deviennent denses dans $\mathcal{X}$ au sens de la norme induite par le produit scalaire. Il résulte de l'équation 1 que la famille $\mathcal{H}$ peut également être considérée comme un modèle affine multivarié sur $\Phi\,(\mathcal{X})$. Les fonctions h prennent alors la forme suivante :

$$h(\cdot) = \left(\langle w_k,\cdot\rangle + b_k\right)_{1\leq k\leq Q},$$

où les vecteurs w_k sont des éléments de $E_{\Phi(\mathcal{X})}$. Ainsi, H_κ apparaît comme l'espace dual de $E_{\Phi(\mathcal{X})}$ (l'espace des formes linéaires sur $E_{\Phi(\mathcal{X})}$). Compte tenu de la définition de $\langle\cdot,\cdot\rangle_{H_\kappa}$, il s'agit même de l'espace de Hilbert dual de $E_{\Phi(\mathcal{X})}$. Dans la suite, pour une fonction h donnée, $\mathbf{w}$ désignera le vecteur $(w_k)_{1\leq k\leq Q}$ de $E^Q_{\Phi(\mathcal{X})}$ et $\mathbf{b}$ le vecteur $(b_k)_{1\leq k\leq Q}$ de $\mathbb{R}^Q$. On notera $\bar{\mathcal{H}}$

l'espace produit H_κ^Q, $\bar{h} = (\langle w_k, \cdot \rangle)_{1 \leq k \leq Q}$ ses fonctions (considérées comme des fonctions sur $\Phi(\mathcal{X})$) et $\| \cdot \|_{\bar{\mathcal{H}}}$ sa norme. On prendra

$$\left\|\bar{h}\right\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^{Q} \left\|\bar{h}_k\right\|_{H_\kappa}^2} = \sqrt{\sum_{k=1}^{Q} \|w_k\|^2},$$

où $\|\cdot\|$ désigne la norme associée au produit scalaire de $E_{\Phi(\mathcal{X})}$.

Pour établir le lien entre une famille de fonctions $\mathcal{H}$ pour $Q = 2$ et la famille de fonctions univariées sur laquelle s'appuie la SVM bi-classe de même noyau, il faut considérer l'hyperplan affine défini par l'équation $\sum_{k=1}^{Q} h_k = 0$. En notant $\tilde{\mathcal{H}}$ la famille des fonctions $\tilde{h}$ réalisables par la SVM, on définit entre $\tilde{\mathcal{H}}$ et l'hyperplan la bijection suivante : $h = \left(\tilde{h}, -\tilde{h}\right)$. Ainsi, pour plonger une SVM bi-classe dans le cadre plus général considéré ici, il suffit de réécrire ses paramètres sous la forme $w_1 = w = -w_2$ et $b_1 = b = -b_2$.

3.1.2 Algorithme d'apprentissage

Dans ce qui suit, Q étant supposé être supérieur ou égal à trois, nous considérons des problèmes de discrimination multi-classe à Q catégories dont $\mathcal{X}$ est l'espace de description et $\mathcal{Y} = [\![1, Q]\!]$ est l'ensemble des catégories. A chaque fonction h de $\mathcal{H}$ est associée une règle de décision f de $\mathcal{X}$ dans $\mathcal{Y} \bigcup \{*\}$, où $*$ représente une catégorie fictive. $f(x) = l$ si et seulement si $h_l(x) > \max_{k \neq l} h_k(x)$. En cas d'ex æquo, $f(x) = *$, ce qui est compté comme une réponse erronée, i.e., contribue au risque. L'introduction de la catégorie fictive se comprend comme un artifice destiné à simplifier l'expression de la règle de décision, et nous négligerons dans la suite son influence sur la consistance des algorithmes d'apprentissage considérés. Ces définitions étant posées, une M-SVM se définit de la manière suivante.

DÉFINITION 1 (M-SVM, Guermeur, 2007b, définition 42) *Soient un "ensemble d'apprentissage" $z^m = ((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times \mathcal{Y})^m$ et $\lambda \in \mathbb{R}_+^*$. Une M-SVM à Q catégories sur $\mathcal{X}$ est un modèle discriminant à grande(s) marge(s) obtenu en minimisant sur l'hyperplan $\sum_{k=1}^{Q} h_k = 0$ de $\mathcal{H}$ une fonction objectif $J_{\text{M-SVM}}$ de la forme :*

$$J_{\text{M-SVM}}(h) = \sum_{i=1}^{m} \ell_{\text{M-SVM}}(y_i, h(x_i)) + \lambda \left\|\bar{h}\right\|_{\bar{\mathcal{H}}}^2$$

où la fonction de perte $\ell_{\text{M-SVM}}$, convexe, est construite autour de la "fonction de perte charnière" $(\cdot)_+$.

Un théorème de représentation établit que la solution optimale de ce problème s'écrit sous la forme suivante :

$$h(\cdot) = \left(\sum_{i=1}^{m} \beta_{ik} \kappa(x_i, \cdot) + b_k \right)_{1 \leq k \leq Q}.$$

3.1.3 Marges géométriques

D'un point de vue géométrique, l'algorithme décrit ci-dessus tend à construire un ensemble de C_Q^2 hyperplans séparateurs maximisant de manière globale les *marges* entre les différentes catégories. Si ces marges sont définies comme dans le cas bi-classe, leur expression analytique est plus complexe.

DÉFINITION 2 (Marges géométriques, Guermeur, 2007a, définition 7) *Considérons une M-SVM à Q catégories (une fonction de $\mathcal{H}$) dont l'erreur sur l'ensemble d'apprentissage z^m est nulle. γ_{kl}, sa* marge relative aux classes d'indices k et *l, est la distance euclidienne minimale (dans l'espace de représentation) entre un point de l'ensemble d'apprentissage dans l'une ou l'autre de ces classes et l'hyperplan les séparant. En notant*

$$d_{\text{M-SVM}} = \min_{1\leq k<l\leq Q} \left\{ \min \left[\min_{i:y_i=k} \left(h_k(x_i) - h_l(x_i)\right), \min_{j:y_j=l} \left(h_l(x_j) - h_k(x_j)\right) \right] \right\}$$

et pour $1 \leq k < l \leq Q$*,*

$$d_{\text{M-SVM},kl} = \frac{1}{d_{\text{M-SVM}}} \min \left[\min_{i:y_i=k} \left(h_k(x_i) - h_l(x_i) - d_{\text{M-SVM}}\right), \min_{j:y_j=l} \left(h_l(x_j) - h_k(x_j) - d_{\text{M-SVM}}\right) \right],$$

l'expression analytique de γ_{kl} est donc :

$$\gamma_{kl} = d_{\text{M-SVM}} \frac{1 + d_{\text{M-SVM},kl}}{\|w_k - w_l\|}.$$

3.2 M-SVM de Weston et Watkins

La M-SVM de Weston et Watkins est l'instanciation du modèle générique décrit ci-dessus dans laquelle la fonction de perte $\ell_{\text{M-SVM}}$ est remplacée par la fonction ℓ_{WW} telle que

$$\ell_{\text{WW}}(y, h(x)) = \sum_{k\neq y} \left(1 - h_y(x) + h_k(x)\right)_+ .$$

En notant C la *constante de marge douce* ($C = (2\lambda)^{-1}$), le problème de programmation quadratique correspondant à son algorithme d'apprentissage est donc le suivant :

PROBLÈME 1 (M-SVM de Weston et Watkins, problème primal)

$$\min_{h\in\mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k\neq y_i} \xi_{ik} \right\}$$

$$\text{s.c.} \left\{ \begin{array}{ll} \langle w_{y_i} - w_k, \Phi(x_i)\rangle + b_{y_i} - b_k \geq 1 - \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \end{array} \right. .$$

Notons en premier lieu que si la contrainte $\sum_{k=1}^{Q} h_k = 0$ n'apparaît pas dans la formulation de ce problème, elle est bien satisfaite par la solution optimale (Guermeur, 2007a, proposition 1). En nous plaçant dans le cadre de l'application de la dualité lagrangienne (Fletcher, 1987), soit α_{ik} le multiplicateur de Lagrange associé à la contrainte de bon classement $h_{y_i}(x_i) - h_k(x_i) \geq 1 - \xi_{ik}$. Afin de rendre l'exposé plus clair, nous utilisons la notation matricielle $\alpha = (\alpha_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$ pour désigner le vecteur de ces multiplicateurs (le passage de la matrice au vecteur s'obtient naturellement par concaténation des vecteurs lignes). Sont introduites à cette occasion, pour i allant de 1 à m, les pseudo-variables α_{iy_i} toutes égales à 0. Ces notations étant posées, le dual de Wolfe du problème 1 prend la forme suivante :

PROBLÈME 2 (M-SVM de Weston et Watkins, problème dual)

$$\min_{\alpha} \left\{ \frac{1}{2} \alpha^T H_{WW} \alpha - 1_{Qm}^T \alpha \right\}$$

$$\textit{s.c.} \begin{cases} 0 \leq \alpha_{ik} \leq C, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{i=1}^{m} \sum_{l=1}^{Q} (\delta_{y_i,k} - \delta_{k,l}) \, \alpha_{il} = 0, & (1 \leq k \leq Q-1) \end{cases},$$

où $H_{\text{WW}} = \left(h_{ik,jl}^{(\text{WW})} \right)_{1 \leq i,j \leq m, 1 \leq k,l \leq Q}$ est la matrice de $\mathcal{M}_{Qm,Qm}(\mathbb{R})$ de terme général

$$h_{ik,jl}^{(\text{WW})} = \left(\delta_{y_i,y_j} - \delta_{y_i,l} - \delta_{y_j,k} + \delta_{k,l} \right) \kappa(x_i, x_j),$$

1_{Qm} est le vecteur de $\mathbb{R}^{Qm}$ dont toutes les composantes sont égales à 1 et δ est le symbole de Kronecker.

3.3 M-SVM de Lee et co-auteurs et M-SVM^2

Pour naturelle qu'apparaisse l'extension de la SVM bi-classe représentée par la M-SVM de Weston et Watkins, elle possède un défaut : la suite des fonctions qu'elle sélectionne ne converge pas vers une fonction dont la règle de décision est le classifieur de Bayes lorsque la taille de l'ensemble d'apprentissage tend vers l'infini (son algorithme d'apprentissage n'est pas *Bayes* ou *Fisher consistant*). Ce résultat peut en particulier être trouvé dans (Zhang, 2004; Tewari et Bartlett, 2007). Le modèle de M-SVM proposé par Lee et co-auteurs est conçu de manière que le principe inductif sur lequel il repose soit universellement consistant.

3.3.1 M-SVM de Lee et co-auteurs

La propriété de consistance est obtenue en choisissant un codage approprié pour les catégories. Il correspond au choix de représentants des catégories situés sur les sommets d'un polytope régulier de $\mathbb{R}^{Q-1}$ à Q sommets centré sur l'origine de $\mathbb{R}^Q$. Plus précisément, la catégorie d'indice k est représentée par le vecteur v_k de $\mathbb{R}^Q$ dont la l-ième composante est égale à 1 si $l = k$ et à $-\frac{1}{Q-1}$ dans le cas contraire. Le polytope en question est donc un simplexe. Le cas où $Q = 3$ est illustré par la figure 2. Naturellement, il peut s'avérer utile de normer les

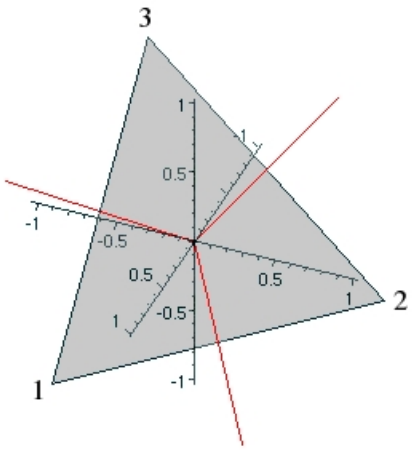

FIG. 2 – *Principe de la M-SVM de Lee et ses co-auteurs illustré dans le cas d'un problème à trois catégories. Les représentants des catégories sont situés sur les sommets d'un triangle équilatéral centré sur l'origine de $\mathbb{R}^3$. Les frontières de décision optimales, qui apparaissent en rouge, sont des demi-droites incluses dans les médianes du triangle et dont l'origine est l'origine de l'espace.*

vecteurs v_k en les multipliant par $\sqrt{\frac{Q-1}{Q}}$. On obtient alors :

$$\forall (k,l) \in [\![1,Q]\!]^2\,, \quad \langle v_k, v_l \rangle = \left(-\frac{1}{Q-1}\right)^{1-\delta_{k,l}}. \tag{2}$$

Cette répartition optimale de points est bien connue en théorie de l'apprentissage. Elle est en particulier utilisée par Vapnik (1982) afin d'établir une variante du lemme de Sauer-Shelah. Elle est également à la base de l'extension multi-classe du principe d'alignement noyau-cible présentée dans la section 5.2. Prolongeant la propriété de sommation à 0 des représentants, les auteurs introduisent, cette fois explicitement, la contrainte $\sum_{k=1}^{Q} h_k = 0$. On voit ainsi apparaître pour fonction de perte

$$\ell_{\text{LLW}}\left(y, h(x)\right) = \sum_{k \neq y} \left(h_k(x) + \frac{1}{Q-1}\right)_+$$

et l'apprentissage consiste donc à résoudre le problème de programmation quadratique suivant :

PROBLÈME 3 (M-SVM de Lee et co-auteurs, problème primal)

$$\min_{h\in\mathcal{H}} \left\{ \frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\sum_{i=1}^{m}\sum_{k\neq y_i} \xi_{ik} \right\}$$

$$\textit{s.c.} \left\{ \begin{array}{ll} \langle w_k, \Phi(x_i)\rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik}, & (1\leq i\leq m), (1\leq k\neq y_i\leq Q) \\ \xi_{ik}\geq 0, & (1\leq i\leq m), (1\leq k\neq y_i\leq Q) \\ \sum_{k=1}^{Q} w_k = 0, \ \ \sum_{k=1}^{Q} b_k = 0 & \end{array} \right. .$$

En notant comme précédemment les multiplicateurs de Lagrange associés aux contraintes de bon classement, on obtient pour problème dual du problème 3 le problème suivant :

PROBLÈME 4 (M-SVM de Lee et co-auteurs, problème dual)

$$\min_{\alpha} \left\{ \frac{1}{2}\alpha^T H_{LLW}\alpha - \frac{1}{Q-1} 1_{Qm}^T \alpha \right\}$$

$$\textit{s.c.} \left\{ \begin{array}{ll} 0\leq \alpha_{ik}\leq C, & (1\leq i\leq m), (1\leq k\neq y_i\leq Q) \\ \sum_{i=1}^{m}\sum_{l=1}^{Q}\left(\frac{1}{Q}-\delta_{k,l}\right)\alpha_{il} = 0, & (1\leq k\leq Q-1) \end{array} \right. ,$$

où la matrice hessienne H_{LLW} a pour terme général

$$h_{ik,jl}^{(LLW)} = \left(\delta_{k,l} - \frac{1}{Q}\right)\kappa(x_i, x_j).$$

3.3.2 M-SVM^2

La SVM bi-classe standard possède une variante dite "de norme 2", pour laquelle la contribution empirique à la fonction objectif est quadratique. Il s'agit précisément de $C\|\xi\|_2^2$. Avec cette variante, la contrainte de positivité des variables d'écart devient superflue (voir en particulier Shawe-Taylor et Cristianini, 2004, chapitre 7). Ce modèle possède une propriété utile : il rend possible, au moyen d'un changement de noyau adéquat, la reformulation de l'algorithme d'apprentissage d'une SVM à marge douce comme l'algorithme d'apprentissage d'une SVM à marge dure. Un avantage immédiat de cette propriété est qu'elle permet l'utilisation de la borne "rayon-marge" (Vapnik, 1998, chapitre 10) afin de déterminer la valeur de la constante de marge douce C. La M-SVM^2 est la généralisation multi-classe de la SVM de norme 2. Il s'agit d'une variante de la M-SVM de Lee et co-auteurs caractérisée par l'expression de la composante empirique de sa fonction objectif. Cette composante est $C\|M\xi\|_2^2$, la matrice $M = (m_{ik,jl})_{1\leq i,j\leq m, 1\leq k,l\leq Q}$ de $\mathcal{M}_{Qm,Qm}(\mathbb{R})$, symétrique, semi-définie positive, ayant pour terme général :

$$m_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q}\right)\delta_{i,j}.$$

M étant idempotente, $\|M\xi\|_2^2 = \xi^T M\xi$. Le problème de programmation quadratique correspondant à l'apprentissage de cette machine est donc le suivant :

PROBLÈME 5 (M-SVM^2, problème primal)

$$\min_{h\in\mathcal{H}} \left\{ \frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\xi^T M\xi \right\}$$

$$s.c. \begin{cases} \langle w_k, \Phi(x_i)\rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{k=1}^{Q} w_k = 0, \ \ \sum_{k=1}^{Q} b_k = 0 \end{cases} .$$

Le dual de Wolfe a pour expression :

PROBLÈME 6 (M-SVM^2, problème dual)

$$\min_{\alpha} \left\{ \frac{1}{2}\alpha^T \left(H_{LLW} + \frac{1}{2C}M \right) \alpha - \frac{1}{Q-1} 1_{Qm}^T \alpha \right\}$$

$$s.c. \begin{cases} \alpha_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq y_i \leq Q) \\ \sum_{i=1}^{m}\sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{il} = 0, & (1 \leq k \leq Q-1) \end{cases} .$$

En comparant ce problème au problème 4, on observe qu'il correspond bien à l'algorithme d'apprentissage d'une M-SVM de Lee et co-auteurs à marges dures (pour laquelle $C = \infty$), le noyau κ étant remplacé par le noyau κ' tel que :

$$\forall (i,j) \in [\![1,m]\!]^2, \ \ \kappa'(x_i, x_j) = \kappa(x_i, x_j) + \frac{1}{2C}\delta_{i,j}.$$

Il est donc possible de déterminer la valeur de la constante de marge douce de la M-SVM^2 en utilisant pour fonction objectif la borne rayon-marge multi-classe suivante :

THÉORÈME 1 (Borne rayon-marge multi-classe, Monfrini et Guermeur, 2008, théorème 2) *Considérons une M-SVM de Lee et co-auteurs à Q catégories à marges dures sur $\mathcal{X}$. Soient z^m son ensemble d'apprentissage, $\mathcal{L}_m$ le nombre d'erreurs résultant de l'application à cette machine d'une procédure de validation croisée* leave-one-out *et $\mathcal{D}_m$ le diamètre de la plus petite sphère de $E_{\Phi(\mathcal{X})}$ contenant l'ensemble des représentations $\{\Phi(x_i) : 1 \leq i \leq m\}$. On dispose alors de la majoration suivante :*

$$\mathcal{L}_m \leq Q^2 \mathcal{D}_m^2 \sum_{k<l} \left(\frac{1 + d_{LLW,kl}}{\gamma_{kl}} \right)^2 .$$

4 Noyau RBF dédié à la prédiction de la structure secondaire

Nous avons vu dans la section 2.2 que la prédiction de la structure secondaire des protéines est un domaine qui est apparu il y a près de quarante ans et a fait depuis lors l'objet de recherches intensives. De nos jours, il n'est plus possible d'espérer dépasser l'état de l'art si ce

n'est en mettant en œuvre un système discriminant spécialement conçu pour cette tâche, système intégrant autant que possible l'importante expertise accumulée au cours des années aussi bien par le biologiste que par le bioinformaticien. Dans le cas où ce système est une méthode à noyau, cela passe naturellement par la spécification d'un nouveau noyau (ou de plusieurs). Celui qui est présenté dans cette section repose sur des considérations biologiques très simples. Il s'agit d'un noyau RBF ou noyau gaussien entièrement caractérisé par le choix de la norme sur l'espace de description (espace des contenus de fenêtres d'analyse). Cette norme est paramétrée, si bien que le travail d'adaptation du noyau aux données relève directement de la sélection de modèle pour les M-SVM, sujet traité dans la section 5.

4.1 Choix des prédicteurs

Ainsi que nous l'avons laissé entendre dans la section 2.2, en évoquant la méthode de prédiction neuronale proposée par Qian et Sejnowski, la manière usuelle d'effectuer la prédiction de la structure secondaire au moyen de modèles de l'apprentissage statistique consiste à employer une approche locale. Dans un but de simplification de l'exposé, nous la présentons dans un premier temps dans le cas le plus simple où les structures primaires faisant l'objet de la prédiction sont utilisées seules, et non incorporées dans un alignement multiple. Les prédicteurs employés pour déterminer l'état conformationnel d'un résidu donné sont alors les acides aminés contenus dans une fenêtre d'analyse de taille fixe centrée sur ce résidu (dans certains cas, la fenêtre est asymétrique). Ce sont ces prédicteurs que nous avons utilisés dans le cadre de notre étude comparative. Dans l'approche classique, un vecteur de 22 composantes est employé pour coder le contenu de chaque position de la fenêtre. Chacune de ses 20 premières composantes est associée à un acide aminé donné (il y en a 20 différents), les deux composantes restantes étant utilisées respectivement pour représenter les acides aminés indéterminés, habituellement représentés par un 'X' dans les bases, et les positions vides de la fenêtre. Une fenêtre d'analyse peut contenir des positions vides lorsqu'elle déborde à l'une des extrémités (N ou C terminale) de la chaîne protéique traitée. En résumé, le codage utilisé pour représenter le contenu de la fenêtre est le codage orthonormal standard, qui n'induit aucune corrélation entre les symboles de l'alphabet. Ce qui apparaît a priori comme un avantage est ici un inconvénient, ainsi que nous le verrons dans la section suivante. Pour une taille de la fenêtre d'analyse $|W| = 2n + 1$, où la valeur de n est habituellement comprise entre 5 et 10, le nombre de prédicteurs est donc égal à $(2n + 1)22$. Seuls $(2n + 1)$ d'entre eux sont égaux à 1, les autres étant égaux à 0. Ceci produit un vecteur de grande taille très creux. La situation est différente lorsqu'un profil d'alignement multiple est utilisé à la place de la structure primaire. Une attention particulière doit être apportée à l'inclusion sous cette forme d'informations évolutives, dans la mesure où elle améliore significativement les performances des méthodes de prédiction, comme nous avons eu l'occasion de l'observer dans la section 2.2. Nous évoquerons également dans la suite l'application de notre noyau à des contenus de fenêtres d'analyse obtenus à partir d'alignements multiples.

4.2 Insuffisances du noyau RBF standard

Considérons le vecteur $\mathbf{x}$ utilisé pour prédire l'état conformationnel d'un résidu donné. Alors, compte tenu du choix des prédicteurs décrit ci-dessus, $\mathbf{x} = (x_i)_{-n \leq i \leq n} \in \{0, 1\}^{(2n+1)22}$, où x_i est le codage canonique de l'acide aminé occupant la position d'indice i dans la fenêtre

d'analyse. En conséquence, l'expression correspondant à l'application d'un noyau RBF standard (utilisant une gaussienne sphérique) à deux contenus de fenêtres $\mathbf{x}$ et $\mathbf{x}'$ se simplifie de la manière suivante :

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) = \exp\left(-\frac{(2n+1) - \sum_{i=-n}^{n} \delta_{x_i, x'_i}}{\sigma^2}\right) \tag{3}$$

où δ est de nouveau le symbole de Kronecker. Le terme de droite de (3) souligne le fait que le noyau dépend uniquement de la distance de Hamming entre les deux segments considérés. Cela constitue un bien pauvre résumé de l'information contenue dans les données. En effet, deux segments correspondant à des parties de protéines homologues superposables dans l'espace, et partageant donc la même structure secondaire, peuvent différer significativement en raison des phénomènes de l'évolution que sont les insertions/délétions ainsi que les substitutions. La distance de Hamming est très sensible aux premiers d'entre eux, tandis qu'elle ne prend pas en compte la nature des substitutions, mais simplement leur nombre. Elle ne prend pas non plus en compte l'influence relative des éléments du contexte en fonction de leur distance au centre de la fenêtre. En conséquence, on ne peut espérer qu'une telle combinaison noyau/codage des données puisse fournir des résultats satisfaisants sur le problème qui nous intéresse. Le constat serait similaire avec les autres noyaux standard. Cette simple observation est à l'origine du travail de développement de noyau décrit dans la section suivante.

4.3 Prise en compte explicite d'informations et connaissances biologiques

L'objectif est ici de prendre en compte dans un noyau RBF deux des facteurs reconnus comme importants pour prédire la structure secondaire : la nature des substitutions entre deux segments (informations évolutives) et l'influence relative des acides aminés impliqués en fonction de leur position dans la fenêtre d'analyse. Chacun de ces points fait l'objet d'une sous-section. Notons que les solutions que nous proposons s'étendent à l'ensemble des noyaux que Williamson et al. (2001) nomment *noyaux de convolution*, c'est-à-dire les noyaux vérifiant $\kappa(t, t') = \kappa(t - t', 0)$.

4.3.1 Produits scalaires entre acides aminés et exploitation d'alignements multiples

Dans la section 4.1, nous avons souligné le fait que le traitement standard des données de séquences en prédiction de la structure secondaire utilise un codage orthonormal des acides aminés. Cependant, il est connu que cette solution n'est pas satisfaisante. De fait, les biologistes ont produit un grand nombre de *matrices de similarité* (on parle également de *matrices de substitution*) pour les acides aminés, qui diffèrent toutes significativement de la matrice identité. C'est en particulier le cas des matrices *percent accepted mutations* (PAM) (Dayhoff et al., 1978) et *blocks substitution matrix* (BLOSUM) (Henikoff et Henikoff, 1992). Le problème soulevé par leur utilisation dans un noyau découle du fait qu'elles ne sont pas symétriques semi-définies positives, et ne sont donc pas associées à un produit scalaire. Différentes solutions peuvent être envisagées pour surmonter cette difficulté. Les matrices étant symétriques, un moyen simple de les approximer par une matrice de Gram consiste à les diagonaliser et à annuler toutes les valeurs propres négatives. La possibilité a priori la meilleure consiste à rechercher leur projection sur l'espace des matrices symétriques semi-définies positives, l'opérateur correspondant étant associé à une norme matricielle, par exemple la norme de Frobenius

(voir la section 5.1). Même s'il se peut que l'on ne dispose pas de l'expression analytique de l'opérateur de projection (le problème à résoudre n'est même pas nécessairement convexe), des estimations satisfaisantes peuvent être obtenues par une simple descente en gradient (Didiot, 2003). Cette descente est alors réalisée par rapport aux composantes de la suite $(a_j)_{1\leq j\leq 22}$ des vecteurs de $\mathbb{R}^{22}$ représentant les différents acides aminés, un contenu indéterminé ou une position vide.

Ce changement de produit scalaire entre acides aminés, qu'il soit issu ou non d'un changement explicite de leur codage, s'étend directement au cas où l'on utilise des alignements multiples (la prise en compte de la présence de *gaps* dans les séquences de ces alignements ne nécessitant pas l'introduction d'un vecteur a_j suplémentaire). Afin de ne pas alourdir les notations, nous illustrons cette extension dans le cas le plus simple où le profil associé à un alignement est obtenu en calculant, pour chaque position de la séquence de base, une moyenne pondérée des vecteurs codant les acides aminés présents en cette position dans au moins l'une des séquences de l'alignement, le poids associé à un acide aminé particulier étant sa fréquence d'occurence dans la position, fréquence mesurée en excluant les *gaps*. Rappelons que ce principe de codage est en particulier celui utilisé par PHD. Notons alors θ_{ij} la fréquence d'apparition de l'acide aminé d'indice j (de codage a_j) dans la position de l'alignement correspondant à la position d'indice i de la fenêtre glissante. Le contenu de la fenêtre peut ainsi être représenté par le vecteur $\tilde{\mathbf{x}} = (\tilde{x}_i)_{-n\leq i\leq n}$ tel que $\tilde{x}_i = \sum_{j=1}^{22} \theta_{ij} a_j$. En conséquence, dans le calcul du noyau, le produit scalaire $\langle x_i, x'_i \rangle$ est simplement remplacé par :

$$\langle \tilde{x}_i, \tilde{x}'_i \rangle = \langle \sum_{j=1}^{22} \theta_{ij} a_j, \sum_{k=1}^{22} \theta'_{ik} a_k \rangle = \sum_{j=1}^{22} \sum_{k=1}^{22} \theta_{ij} \theta'_{ik} \langle a_j, a_k \rangle. \tag{4}$$

On tire ici profit de la linéarité du produit scalaire. Naturellement, comme dans le cas du *kernel trick*, l'écriture $\langle a_j, a_k \rangle$ n'implique pas que l'on dispose de l'expression explicite de $(a_j)_{1\leq j\leq 22}$. Dans la suite, par défaut, les formules sont exprimées dans le cas où les données d'entrée sont des alignements multiples. Le produit scalaire entre les contenus des positions de mêmes indices de deux fenêtres d'analyse n'est alors pas nécessairement donné par l'équation 4, ce qui permet en particulier d'inclure le cas où les profils sont issus de l'algorithme PSI-BLAST.

4.3.2 Influence de la position dans la fenêtre

Nous avons vu que l'utilisation d'une fenêtre glissante était la norme en prédiction de la structure secondaire des protéines. De nombreuses études ont porté sur le choix de sa taille ou l'exploitation de son contenu. De bonnes illustrations sont fournies par Qian et Sejnowski (1988); Zhang et al. (1992); Rost et Sander (1993); Hua et Sun (2001). En bref, une fenêtre trop petite ne contiendra pas assez d'information sur la conformation locale, tandis qu'une fenêtre trop grande incorporera des données qui risquent de se comporter comme du bruit. Un moyen de surmonter cette difficulté consiste à choisir a priori une valeur élevée pour la taille de la fenêtre et à inférer empiriquement une pondération associant à chaque position un coefficient (indépendant de la nature de l'acide aminé), de manière à moduler son influence sur les calculs ultérieurs. La procédure peut être appliquée soit dans le cadre d'une prédiction directe des trois catégories, soit dans le cadre d'une décomposition un contre tous. Dans le second cas, il est instructif de comparer les pondérations obtenues pour les différents états

conformationnels. Un tel apprentissage avait déjà été réalisé avec succès par différentes équipes (Gascuel et Golmard, 1988; Guermeur, 1997). Un point important est que ces études, bien qu'elles aient été fondées sur des approches très différentes, ont produit des distributions des poids très similaires. Ceci suggère qu'elles ont permis de mettre en évidence une propriété intrinsèque du problème considéré. Nous avons donc décidé d'incorporer une telle pondération dans notre noyau.

4.4 Expression analytique du noyau

La combinaison des deux paramétrisations : la modification du "produit scalaire entre acides aminés" et la pondération des positions de la fenêtre d'analyse, conduit à la spécification du noyau suivant.

DÉFINITION 3 (Noyau RBF dédié à la prédiction de la structure secondaire) *$\tilde{\mathbf{x}}$ et $\tilde{\mathbf{x}}'$ étant les contenus de deux fenêtres d'analyse de taille $2n+1$ sur des alignements multiples, $\theta = (\theta_i)_{-n \leq i \leq n}$ un vecteur de poids et $D \in \mathcal{M}_{22,22}(\mathbb{R})$ une matrice de produits scalaires entre acides aminés ($D = (d_{jk})_{1 \leq j,k \leq 22}$ avec $d_{jk} = \langle a_j, a_k \rangle$), le noyau $\kappa_{\theta,D}$ est défini par :*

$$\kappa_{\theta,D}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = \exp\left(- \sum_{i=-n}^{n} \theta_i^2 \left(\|\tilde{x}_i\|^2 + \|\tilde{x}'_i\|^2 - 2\langle \tilde{x}_i, \tilde{x}'_i \rangle \right) \right). \tag{5}$$

5 Sélection de modèle

Pour les M-SVM qui nous intéressent, le problème de la sélection de modèle porte sur le choix de la valeur de la constante de marge douce et des valeurs des paramètres du noyau. Pour fixer la valeur de C, la solution la plus naturelle consiste à appliquer une procédure de validation croisée. Cependant, il est également possible d'utiliser une borne sur le risque, telle que la borne VC introduite dans (Guermeur, 2007b) ou celle reposant sur la complexité de Rademacher (Guermeur, 2007a, théorème 9). Naturellement, dans le cas particulier de la M-SVM^2, l'emploi de la borne rayon-marge multi-classe (théorème 1) représente également une possibilité. Nous avons vu dans la section 4.3.1 qu'il existait au moins deux moyens simples de dériver une matrice de produits scalaires entre acides aminés à partir d'une matrice de substitution. Dans cette section, nous nous focalisons donc sur le calcul de la pondération des positions de la fenêtre d'analyse. Pour effectuer cette tâche, nous proposons d'employer une extension multi-classe du principe d'alignement noyau-cible.

5.1 Alignement de noyaux

L'alignement de noyaux a été introduit par Cristianini et al. (2002), comme un moyen d'évaluer le degré d'adéquation d'un noyau à une tâche d'apprentissage donnée, et adapter en conséquence la matrice de Gram afin d'augmenter cette adéquation. Il s'agit donc fondamentalement d'une méthode conçue pour la transduction (voir en particulier Lanckriet et al., 2004; Vapnik, 1998, chapitre 8) dans la mesure où elle ne fournit pas d'expression analytique pour le noyau résultant.

DÉFINITION 4 (Alignement de noyaux, Cristianini et al., 2002) *Soient κ et κ' deux fonctions noyau mesurables définies sur $\mathcal{T} \times \mathcal{T}$ où $\mathcal{T}$ est supposé être un espace probabilisé muni d'une mesure de probabilité $P_{\mathcal{T}}$. L'*alignement entre κ et κ'*, $A(\kappa, \kappa')$, est défini comme suit :*

$$A(\kappa, \kappa') = \frac{\langle \kappa, \kappa' \rangle}{\|\kappa\| \|\kappa'\|} = \frac{\int_{\mathcal{T}^2} \kappa(t,t')\kappa'(t,t') dP_{\mathcal{T}}(t) dP_{\mathcal{T}}(t')}{\sqrt{\int_{\mathcal{T}^2} \kappa(t,t')^2 dP_{\mathcal{T}}(t) dP_{\mathcal{T}}(t')} \sqrt{\int_{\mathcal{T}^2} \kappa'(t,t')^2 dP_{\mathcal{T}}(t) dP_{\mathcal{T}}(t')}}.$$

DÉFINITION 5 (Alignement empirique de noyaux, Cristianini et al., 2002, définition 1) *$\mathcal{T}$, κ et κ' étant définis comme dans la définition 4, soit $T^n = (T_i)_{1 \leq i \leq n}$ un n-échantillon de variables aléatoires indépendantes distribuées suivant $P_{\mathcal{T}}$. L'alignement de κ et κ' par rapport à T^n est la quantité :*

$$\hat{A}_{T^n}(G, G') = \frac{\langle G, G' \rangle_F}{\|G\|_F \|G'\|_F} \tag{6}$$

où G et G' sont les matrices de Gram associées respectivement à κ et κ', calculées sur T^n, et $\langle \cdot, \cdot \rangle_F$ représente le produit scalaire de Frobenius entre matrices, si bien que $\langle G, G' \rangle_F = \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(T_i, T_j) \kappa'(T_i, T_j)$. $\| \cdot \|_F$ représente la norme correspondante.

L'alignement de noyaux est donc un cosinus, et comme tel fournit une mesure de la similarité des deux vecteurs unitaires concernés. Si κ est un noyau bien adapté au problème considéré et que κ' est bien aligné avec κ, alors κ' est également un bon noyau pour le même problème. En pratique, l'alignement n'étant pas calculable, puisque la distribution sous-jacente $P_{\mathcal{T}}$ est inconnue, il est estimé empiriquement au moyen de la formule 6. Cristianini et ses co-auteurs ont étudié les propriétés de concentration de la variable aléatoire $\hat{A}_{T^n}(G, G')$ autour de son espérance $A(\kappa, \kappa')$.

5.2 Alignement noyau-cible multi-classe : application au paramétrage d'un noyau

Considérons une famille de noyaux κ_θ de paramètre formel θ appartenant à l'ensemble Θ. L'alignement de noyaux étant défini, notre stratégie pour l'appliquer à la détermination partielle ou entière de θ peut être résumée de la manière suivante :

1. choisir un noyau théoriquement idéal κ_t, que nous appellerons dans la suite le *noyau cible*, idéal au sens où il conduit à un classement parfait (en pratique, la matrice de Gram de κ_t doit pouvoir être calculée) ;
2. étant donné un ensemble d'apprentissage $z^m = ((x_i, y_i))_{1 \leq i \leq m}$, choisir θ^* en application du critère suivant :
$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \hat{A}_{z^m}(G_\theta, G_t),$$
où G_θ est la matrice de Gram associée au couple (κ_θ, z^m), G_t étant la matrice de Gram associée au couple (κ_t, z^m).

La conjecture qui est faite est que le noyau κ_{θ^*} ainsi choisi se comportera bien sur le problème considéré, pourvu que la famille $\{\kappa_\theta : \theta \in \Theta\}$ soit appropriée. Dans le cas bi-classe (pour lequel $\mathcal{Y} = \{-1, 1\}$), le noyau idéal se définit de la manière suivante : $\forall ((x, y), (x', y')) \in (\mathcal{X} \times \mathcal{Y})^2$, $\kappa_t(x, x') = yy'$. L'extension multi-classe proposée par Régis Vert (2002) reprend

le raisonnement géométrique sous-jacent, en utilisant pour représentants des catégories les sommets du simplexe décrit dans la section 3.3.1. Compte tenu de l'équation 2, on obtient ainsi une formulation initiale du noyau cible dans le cas de Q catégories :

$$\kappa_t(x, x') = \left(-\frac{1}{Q-1}\right)^{1-\delta_{y,y'}}.$$

Un changement de base, préservant les propriétés géométriques d'intérêt, permet alors de remplacer cette formulation par la suivante, plus simple :

$$\kappa_t(x, x') = \delta_{y,y'}.$$

Notons que sous certaines hypothèses de régularité sur κ_θ, $\hat{A}_{z^m}(G_\theta, G_t)$ est différentiable par rapport à θ, et peut donc être optimisé en utilisant des techniques classiques, telles que la descente en gradient. Nous verrons dans la section 6.2 que c'est le cas avec notre noyau.

6 Evaluation des performances des M-SVM utilisant le noyau RBF dédié

Nous avons souligné dans l'état de l'art que les meilleures méthodes de prédiction sont des systèmes très complexes, parfois constitués de centaines de modules. L'objet de cet article n'est pas d'introduire une nouvelle méthode permettant de les concurrencer. Il ne s'agit que d'une étude préparatoire à la mise au point d'une telle méthode. Les expériences relatées ci-dessous ont pour seule ambition de comparer les performances des M-SVM présentées dans la section 3 munies du noyau décrit dans la section 4, le choix des valeurs des hyper-paramètres étant réalisé au moyen des méthodes décrites dans la section 5. L'étalon utilisé pour cette comparaison est l'unité de base de la plupart des méthodes constituant l'état de l'art, le PMC. Le raisonnement est simple : si les M-SVM s'avèrent supérieures à cette unité de base, alors leur utilisation peut ouvrir la voie à une amélioration de la qualité des prédictions.

6.1 Choix des données et critères d'évaluation des performances

Pour évaluer nos classifieurs, nous avons utilisé l'ensemble de 1096 protéines introduit dans (Guermeur et al., 2004) sous le nom P1096. Cet ensemble a été constitué par Gianluca Pollastri de manière à satisfaire les exigences les plus fortes en termes de taux d'identité (voir Sander et Schneider, 1991, pour les détails). L'assignation de la structure secondaire a été effectuée au moyen du programme DSSP. La procédure utilisée pour réaliser le regroupement des huit états conformationnels initiaux dans les trois catégories de base est celle de CASP : H+G $\rightarrow$ H (hélice α), E+B $\rightarrow$ E (brin β), tous les autres états étant associés à la catégorie C (structure apériodique). Cette façon de procéder, nommée méthode A par Cuff et Barton (1999), est connue comme étant celle conduisant au problème de prédiction le plus difficile.

En présentant le problème de la prédiction de la structure secondaire, nous avons expliqué que le but premier du biologiste est de connaître la structure tridimensionnelle des protéines, afin de tenter d'inférer leur fonction. De ce fait, celui-ci souhaite principalement être en mesure d'identifier l'ensemble des éléments structuraux, avec leur ordre d'apparition sur la séquence.

Un petit décalage entre les positions réelles et prédites des structures peut être toléré, mais la prédiction doit demeurer biologiquement plausible : aucune hélice ne doit être constituée de moins de quatre résidus (un tour d'hélice correspond à 3,6 résidus), deux structures périodiques ne peuvent pas être consécutives... Par suite, le seul taux de reconnaissance par résidu, noté Q_3 dans la littérature, n'est pas suffisant pour caractériser la qualité de la prédiction. De nombreuses mesures de qualité ont été introduites afin de rendre l'évaluation plus fine. Le lecteur intéressé trouvera dans (Baldi et al., 2000) un exposé de synthèse sur la question. Dans ce qui suit, nous utilisons trois des mesures de qualité les plus usuelles : Q_3, les coefficients de corrélation de Pearson/Matthews (Matthews, 1975) et les coefficients Sov déjà évoqués dans la section 2.2. Ils fournissent des informations complémentaires. Tandis que les coefficients de Matthews caractérisent la qualité de la prédiction pour chaque état conformationnel (H/E/C) pris individuellement, ce qui permet, par exemple, de souligner une mauvaise reconnaissance des feuillets, la valeur des coefficients Sov donne une idée de la qualité de la prédiction au niveau des segments, satisfaisant ainsi l'une des principales exigences évoquées au début du paragraphe.

6.2 Estimation des hyper-paramètres

La matrice des produits scalaires entre acides aminés a été obtenue à partir de la matrice de similarité introduite par Levin et al. (1986). La raison de ce choix tient au fait que cette matrice a été spécifiquement conçue pour effectuer la prédiction de la structure secondaire en s'appuyant sur la similarité de petits peptides, c'est-à-dire l'homologie de séquence locale (voir aussi Levin et Garnier, 1988; Geourjon et Deléage, 1995). Dans ce contexte, elle s'est avérée supérieure à la matrice PAM évoquée dans la section 4.3.1. Parmi les deux options considérées dans cette même section pour engendrer la matrice de Gram, nous avons retenu celle utilisant la diagonalisation. Cependant, ce choix a en premier lieu été effectué dans un souci de reproductibilité des expériences, dans la mesure où la descente en gradient fournit des résultats très similaires (Didiot, 2003).

Pour ces valeurs des produits scalaires entre acides aminés et une taille de la fenêtre d'analyse de treize ($n = 6$), la valeur du vecteur de poids θ a été obtenue par application du principe d'alignement noyau-cible multi-classe, au moyen d'une descente en gradient stochastique (à chaque pas de la descente, l'optimisation était réalisée par rapport à un nouveau sous-ensemble de la base d'apprentissage choisi aléatoirement suivant une distribution uniforme). L'expression analytique du gradient est particulièrement simple. En notant $G_{\theta,D}$ la matrice de Gram associée au noyau $\kappa_{\theta,D}$ et $G'_{\theta_k,D}$ la matrice de $\mathcal{M}_{m,m}(\mathbb{R})$ de terme général $\frac{\partial}{\partial\theta_k}\kappa_{\theta,D}\left(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j\right)$, on obtient en effet :

$$\forall k \in [\![-n,n]\!],\ \ \frac{\partial}{\partial\theta_k}\hat{A}_{z^m}(G_{\theta,D},G_t) = \frac{\left\langle G'_{\theta_k,D},G_t\right\rangle_F}{\|G_{\theta,D}\|_F\|G_t\|_F} - \frac{\left\langle G_{\theta,D},G_t\right\rangle_F\left\langle G_{\theta,D},G'_{\theta_k,D}\right\rangle_F}{\|G_{\theta,D}\|_F^3\|G_t\|_F}.$$

La base d'apprentissage utilisée pour l'estimation de ce vecteur était constituée des 1180 séquences employées pour entraîner les méthodes de prédiction SSpro1 et SSpro2 (Baldi et al., 1999; Pollastri et al., 2002). Cet ensemble est nommé dans la suite P1180. Ce choix était possible dans la mesure où aucune séquence de cette base ne présente avec une séquence de la base P1096 un taux d'identité supérieur à 25% (voir aussi Guermeur et al., 2004). En fait, la base P1096 a été assemblée en prenant en compte cette contrainte. La figure 3 illustre les

valeurs obtenues pour les coefficients θ_i lorsque le noyau $\kappa_{\theta,D}$ exploite les structures primaires seules et non des alignements multiples. Cette courbe est très similaire à celles évoquées dans

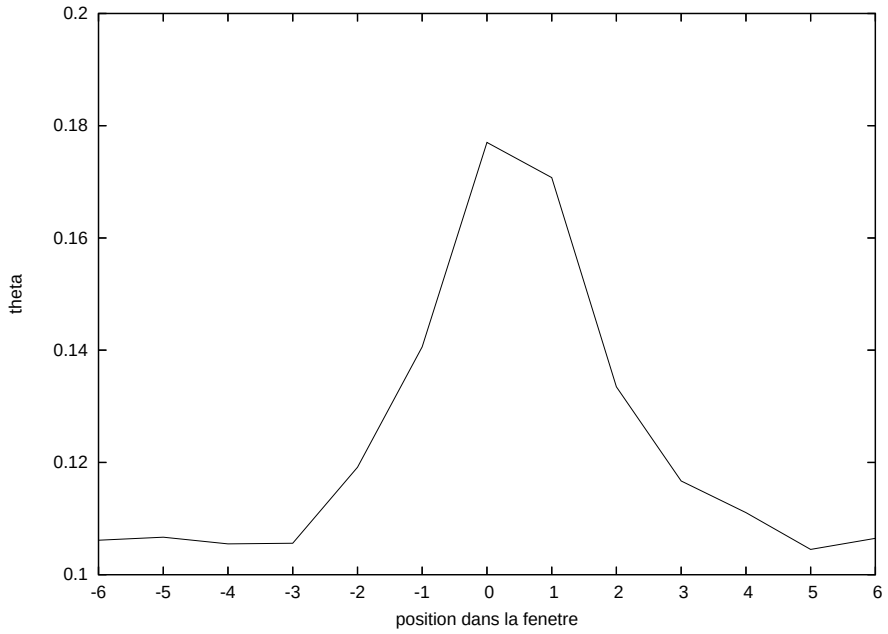

FIG. 3 – *Vecteur θ du noyau $\kappa_{\theta,D}$ défini par l'équation 5 avec $n = 6$, optimisé par alignement noyau-cible multi-classe.*

la section 4.3.2. L'une de leurs caractéristiques communes est leur asymétrie marquée en faveur du contexte droit (du côté de l'extrémité C terminale de la séquence). A notre connaissance, ce phénomène n'a pas trouvé à ce jour d'explication biologique.

Les solutions disponibles pour déterminer la valeur de la constante de marge douce en employant la base déjà utilisée pour entraîner les M-SVM (i.e., P1096), ceci sans rencontrer de problème dû au biais, sont celles évoquées en introduction de la section 5. La procédure de validation croisée appropriée est la *stacked generalization* (Wolpert, 1992). Son principal défaut étant sa complexité (elle correspond à deux procédures de validation croisée imbriquées), nous lui avons préféré l'utilisation de bornes. Pour la M-SVM de Weston et Watkins et celle de Lee et co-auteurs, le choix qui s'imposait était la borne reposant sur la complexité de Rademacher. A l'inverse, pour la M-SVM^2, nous avions le choix entre cette borne et la borne rayon-marge. Des essais préliminaires ayant permis de constater que pour cette machine, l'utilisation des deux bornes produisait des résultats qui ne différaient pas de manière statistiquement significative, dans un but de simplification, la section suivante ne fournit dans tous les cas des performances que pour la première d'entre elles.

6.3 Expérience effectuée et résultats obtenus

Les trois M-SVM et un PMC ont été appliqués aux vecteurs de prédicteurs $\mathbf{x}$ correspond simplement au contenu d'une fenêtre d'analyse (de taille treize) glissant avec un pas d'un résidu sur les structures primaires des protéines. La procédure expérimentale était une validation croisée à cinq pas. Le PMC possédait une couche cachée de dix unités munies d'une sigmoïde. La fonction d'activation des unités de sortie était également une sigmoïde, les essais effectués avec la fonction softmax ayant produit des résultats inférieurs. Les valeurs obtenues pour la constante de marge douce étaient respectivement de un pour la M-SVM de Weston et Watkins et trois pour les deux autres machines. Les résultats observés sont résumés dans le tableau 1.

	PMC	M-SVM WW	M-SVM LLW	M-SVM^2
Q_3	66,0	66,9	66,7	66,7
C_α	0,50	0,52	0,51	0,51
C_β	0,41	0,42	0,40	0,41
C_c	0,45	0,46	0,46	0,46
Sov	55,7	56,0	56,2	56,1
Sov_α	57,7	59,5	62,2	60,1
Sov_β	49,4	51,7	46,7	51,2
Sov_c	57,8	58,4	58,7	58,0

TAB. 1 – *Performances relatives d'un PMC et de trois M-SVM mesurées sur la base P1096.*

Compte tenu de la taille de la base utilisée, 1096 séquences constituées de 268575 résidus, le gain de performance résultant du remplacement du PMC par l'une des M-SVM apparaît statistiquement significatif avec une confiance excédant 0,95. A l'inverse, le test de moyenne standard ne permet pas de distinguer les différentes M-SVM. Dans le détail, les comportements de ces machines sont toutefois loin d'être identiques. Afin de mettre en évidence ce phénomène, on peut en particulier noter que la M-SVM de Weston et Watkins obtient pour C égal à un un taux de reconnaissance en apprentissage de 75,5%, tandis que pour C égal à trois, la M-SVM de Lee et co-auteurs produit un taux de reconnaissance en apprentissage de 73,4%. A titre de comparaison, le taux de reconnaissance en apprentissage du PMC était de seulement 67,9%. Pour ce réseau de neurones, le phénomène de sur-apprentissage survenait beaucoup plus tôt (augmenter la taille de la couche cachée au-delà de dix détériorait les performances en test). Ces dernières remarques soulignent l'importance centrale du problème de la sélection de modèle en prédiction de la structure secondaire. Elles fournissent également de nouveaux arguments en faveur d'une prédiction s'appuyant sur l'emploi d'une ou plusieurs méthodes d'ensemble.

7 Conclusions et perspectives

Nous avons comparé les performances de la M-SVM de Weston et Watkins, la M-SVM de Lee et co-auteurs et la M-SVM^2 en prédiction de la structure secondaire des protéines. Ces performances apparaissent très similaires, difficiles à distinguer à la fois d'un point de vue statistique et d'un point de vue biologique. A l'inverse, la différence entre le plus petit des taux de reconnaissance de ces machines et le taux de reconnaissance d'un PMC s'avère être

statistiquement significative avec une confiance excédant 0,95. Cette étude, certes préliminaire, se revèle donc suffisante pour fournir une première idée du potentiel que peut offrir l'utilisation de M-SVM en prédiction de la structure secondaire.

Naturellement, une alternative à la sélection de modèle est la combinaison de modèles. Comme dans le cas de toutes les familles de classifieurs, son intérêt est d'autant plus important que les erreurs des modèles sont moins corrélées. L'étude de cette corrélation, et au-delà celle de la meilleure méthode pour combiner les sorties de M-SVM, font l'objet d'un travail en cours. Ce travail porte également sur l'exploitation optimale des alignements multiples.

Remerciements

Nous souhaitons exprimer notre profonde gratitude à Gianluca Pollastri pour avoir mis à notre disposition une version actualisée de sa base P1096. C'est également un plaisir de remercier Fabienne Thomarat, Alain Lifchitz et Christophe Magnan pour leurs relectures attentives de cet article.

Références

Altschul, S. et E. Koonin (1998). Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. *Trends in Biochemical Sciences 23*(11), 444–447.

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

Anfinsen, C., E. Haber, M. Sela, et F. White (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences 47*(9), 1309–1314.

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

Asai, K., S. Hayamizu, et K. Handa (1993). Prediction of protein secondary structure by the hidden Markov model. *CABIOS 9*(2), 141–146.

Baldi, P., S. Brunak, Y. Chauvin, C. Andersen, et H. Nielsen (2000). Assessing the accuracy of prediction algorithms for classification : an overview. *Bioinformatics 16*(5), 412–424.

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

Berlinet, A. et C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston.

Berman, H., J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, et P. Bourne (2000). The protein data bank. *Nucleic Acids Research 28*(1), 235–242.

Bernstein, F., T. Koetzle, G. Williams, E. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, et M. Tasumi (1977). The protein data bank : a computer-based archival file for macromolecular structures. *Journal of Molecular Biology 112*(3), 535–542.

Boscott, P., G. Barton, et W. Richards (1993). Secondary structure prediction for modelling by homology. *Protein Engineering 6*(3), 261–266.

Chou, P. et G. Fasman (1978). Empirical predictions of protein conformation. *Annual Review of Biochemistry 47*, 251–276.

Combet, C., M. Jambon, G. Deléage, et C. Geourjon (2002). Geno3D : automatic comparative molecular modelling of protein. *Bioinformatics 18*(1), 213–214.

Crammer, K. et Y. Singer (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research 2*, 265–292.

Cristianini, N., J. Shawe-Taylor, A. Elisseeff, et J. Kandola (2002). On kernel-target alignment. In *NIPS 14*, pp. 367–373.

Cuff, J. et G. Barton (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins : Structure, Function, and Genetics 34*(4), 508–519.

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

Dayhoff, M., R. Schwartz, et B. Orcutt (1978). A model of evolutionary change in proteins. In M. Dayhoff (Ed.), *Atlas of Protein Sequence and Structure*, Volume 5, pp. 345–352. Silver Spring, Washington DC : National Biomedical Research Foundation.

Derrick, J. et D. Wigley (1994). The third IgG-binding domain from streptococcal protein G : An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *Journal of Molecular Biology 243*(5), 906–918.

Didiot, E. (2003). Conception et mise en œuvre de M-SVM dédiées au traitement de séquences biologiques. Mémoire de DEA, DEA informatique de Lorraine.

Dodge, C., R. Schneider, et C. Sander (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Research 26*(1), 313–315.

Epstein, C., R. Goldberger, et C. Anfinsen (1963). The genetic control of tertiary protein structure : Studies with model systems. In *Cold Spring Harbor Symposium on Quantitative Biology*, Volume 28, pp. 439–449.

Fletcher, R. (1987). *Practical Methods of Optimization* (Second ed.). John Wiley & Sons, Chichester.

Gaboriaud, C., V. Bissery, T. Benchetrit, et J.-P. Mornon (1987). Hydrophobic cluster analysis : an efficient new way to compare and analyse amino acid sequences. *FEBS letters 224*(1), 149–155.

Gascuel, O. et J.-L. Golmard (1988). A simple method for predicting the secondary structure of globular proteins : implications and accuracy. *CABIOS 4*(3), 357–365.

Geourjon, C. et G. Deléage (1995). SOPMA : significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS 11*(6), 681–684.

Gibrat, J.-F., J. Garnier, et B. Robson (1987). Further developments of protein secondary structure prediction using information theory. *Journal of Molecular Biology 198*(3), 425–

443.

Guermeur, Y. (1997). *Combinaison de Classifieurs Statistiques, Application à la Prédiction de la Structure Secondaire des Protéines*. Thèse de doctorat, Université Paris 6.

Guermeur, Y. (2000). Combining discriminant models with new multi-class SVMs. Technical Report NC2-TR-2000-086, NeuroCOLT2.

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

Guermeur, Y. (2007a). *SVM multiclasses, théorie et applications*. Habilitation à diriger des recherches, Université Nancy 1.

Guermeur, Y. (2007b). VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research 8*, 2551–2594.

Guermeur, Y., G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy, et P. Baldi (2004). Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing 56C*, 305–327.

Guo, J., H. Chen, Z. Sun, et Y. Lin (2004). A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins : Structure, Function, and Bioinformatics 54*(4), 738–743.

Hardin, C., T. Pogorelov, et Z. Luthey-Schulten (2002). *Ab initio* protein structure prediction. *Current Opinion in Structural Biology 12*(2), 176–181.

Henikoff, S. et J. Henikoff (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America 89*(22), 10915–10919.

Hua, S. et Z. Sun (2001). A novel method of protein secondary structure prediction with high segment overlap measure : Support vector machine approach. *Journal of Molecular Biology 308*(2), 397–407.

Jones, D. (1997). Successful *ab initio* prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins : Structure, Function, and Genetics S1*, 185–191.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology 292*(2), 195–202.

Kabsch, W. et C. Sander (1983). Dictionary of protein secondary structure : Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*(12), 2577–2637.

Karplus, M. et G. Petsko (1990). Molecular dynamics simulations in biology. *Nature (London) 347*, 631–639.

Kim, H. et H. Park (2003). Protein secondary structure prediction based on an improved support vector machine approach. *Protein Engineering 16*(8), 553–560.

Lanckriet, G., N. Cristianini, P. Bartlett, L. El Ghaoui, et M. Jordan (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research 5*, 27–72.

Lee, Y., Y. Lin, et G. Wahba (2004). 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.

Lemer, C., M. Rooman, et S. Wodak (1995). Protein structure prediction by threading methods : evaluation of current techniques. *Proteins : Structure, Function, and Genetics 23*(3), 337–355.

Levin, J. et J. Garnier (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochimica et Biophysica Acta 955*(3), 283–295.

Levin, J., B. Robson, et J. Garnier (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS 205*(2), 303–308.

Lim, V. (1974a). Algorithms for prediction of α-helical and β-structural regions in globular proteins. *Journal of Molecular Biology 88*(4), 873–894.

Lim, V. (1974b). Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *Journal of Molecular Biology 88*(4), 857–872.

Marin, A., J. Pothier, K. Zimmermann, et J.-F. Gibrat (2002). FROST : a filter-based fold recognition method. *Proteins : Structure, Function, and Genetics 49*(4), 493–509.

Martin, J. (2005). *Prédiction de la structure locale des protéines par des modèles de chaînes de Markov cachées*. Thèse de doctorat, Université Paris 7.

Martin, J., J.-F. Gibrat, et F. Rodolphe (2005). Choosing the optimal hidden Markov model for secondary-structure prediction. *IEEE Intelligent Systems 20*(6), 19–25.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta 405*(2), 442–451.

Meiler, J. et D. Baker (2003). Coupled prediction of protein secondary and tertiary structure. *Procedings of the National Academy of Sciences of the United States of America 100*(21), 12105–12110.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London, Series A 209*, 415–446.

Monfrini, E. et Y. Guermeur (2008). A quadratic loss multi-class SVM. Technical report, LORIA, hal-00276700.

Nguyen, M. et J. Rajapakse (2003). Multi-class support vector machines for protein secondary structure prediction. *Genome Informatics 14*, 218–227.

Nguyen, M. et J. Rajapakse (2005). Two-stage multi-class support vector machines to protein secondary structure prediction. In *Pacific Symposium on Biocomputing 10*, pp. 346–357.

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

Pollastri, G., D. Przybylski, B. Rost, et P. Baldi (2002). 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.

Qian, N. et T. Sejnowski (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology 202*(4), 865–884.

Rifkin, R. et A. Klautau (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research 5*, 101–141.

Riis, S. et A. Krogh (1996). Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *Journal of Computational Biology 3*, 163–183.

Rost, B. (2001). Review : Protein secondary structure prediction continues to rise. *Journal of Structural Biology 134*(2), 204–218.

Rost, B. et C. Sander (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology 232*(2), 584–599.

Rost, B. et C. Sander (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins : Structure, Function, and Genetics 19*(1), 55–72.

Rost, B., C. Sander, et R. Schneider (1994). Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology 235*(1), 13–26.

Rost, B., R. Schneider, et C. Sander (1997). Protein fold recognition by prediction-based threading. *Journal of Molecular Biology 270*(3), 471–480.

Russell, R., R. Copley, et G. Barton (1996). Protein fold recognition by mapping predicted secondary structures. *Journal of Molecular Biology 259*(3), 349–365.

Sali, A. (1995). Modelling mutations and homologous proteins. *Current Opinion in Biotechnology 6*(4), 437–451.

Sander, C. et R. Schneider (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins : Structure, Function, and Genetics 9*(1), 56–68.

Sayle, R. et E. Milner-White (1995). RasMol : Biomolecular graphics for all. *Trends in Biochemical Sciences 20*(9), 374–376.

Shawe-Taylor, J. et N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Simons, K., R. Bonneau, I. Ruczinski, et D. Baker (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins : Structure, Function, and Genetics Suppl. 3*, 171–176.

Solovyev, V. et A. Salamov (1994). Predicting α-helix and β-strand segments of globular proteins. *CABIOS 10*(6), 661–669.

Tewari, A. et P. Bartlett (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research 8*, 1007–1025.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons, Inc., New York.

Vert, R. (2002). Conception et mise en œuvre de M-SVM dédiées au traitement de séquences biologiques. Mémoire de DEA, DEA informatique de Lorraine.

Wang, L.-H., J. Liu, Y.-F. Li, et H.-B. Zhou (2004). Predicting protein secondary structure by a support vector machine based on a new coding scheme. *Genome Informatics 15*(2),

181–190.

Ward, J., L. McGuffin, B. Buxton, et D. Jones (2003). Secondary structure prediction with support vector machines. *Bioinformatics 19*(13), 1650–1655.

Weston, J. et C. Watkins (1998). Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science.

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

Wolpert, D. (1992). Stacked generalization. *Neural Networks 5*(2), 241–259.

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

Zhang, Q., S. Yoon, et W. Welsh (2005). Improved method for predicting β-turn using support vector machine. *Bioinformatics 21*(10), 2370–2374.

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

Zhang, X., J. Mesirov, et D. Waltz (1992). Hybrid system for protein secondary structure prediction. *Journal of Molecular Biology 225*(4), 1049–1063.

Summary

Bi-class SVMs, introduced in bioinformatics at the end of the nineties, currently obtain state-of-the-art performance for numerous problems of biological sequence processing. Multi-class SVMs, of more recent conception, are progressively applied to these problems, especially in predictive structural biology. This article deals with a comparative study of the performance of three multi-class SVMs in protein secondary structure prediction. The models involved are the one of Weston and Watkins, the one of Lee and co-authors, and a new machine named M-SVM^2. This study must be considered as a step in the design of a hybrid prediction method, integrating discriminant and generative models, and based on a hierarchical approach of the problem.