# Un noyau d'alignement local pour la classification de séquences biologiques

**Jean-Philippe Vert*** — **Hiroto Saigo**** — **Tatsuya Akutsu****

** Ecole des Mines de Paris*
*Centre de Géostatistique*
*35 rue Saint-Honoré*
*77300 Fontainebleau*
*Jean-Philippe.Vert@ensmp.fr*
*** Kyoto University*
*Institute for Chemical Research*
*Bioinformatics Center*
*Uji 611-0011, Japan*
*{hiroto,takutsu}@kuicr.kyoto-u.ac.jp*

RÉSUMÉ. *La recherche d'alignements entre séquences biologiques est un outil couramment utilisé pour la recherche d'homologie et donc de similarité entre séquences. Nous montrons comment ce concept peut être adapté à des problèmes de classification supervisée de séquences biologiques à l'aide de machines à vecteurs de support, ou de toute autre méthode utilisant des noyaux positifs. Pour ce faire nous montrons comment un noyau défini positif peut être construit à partir du score d'alignement local utilisé pour la recherche d'homologie, à l'aide d'opérations de convolution entre noyaux. Des expériences de classification de séquences en super-familles structurelles valident cette approche.*

MOTS-CLÉS : *Séquences biologiques, alignement local, noyau défini positif, machine à vecteurs de support*

## 1. Introduction

Alors que les quantités de séquences biologiques générées par les différents programmes de séquençages continuent de croître à grande vitesse, les besoins en algorithmes performants pour analyser et classer ces séquences se font de plus en plus pressants. En particulier, la classification automatique de gènes en classes structurelles ou fonctionnelles est un pré-requis pour la compréhension des fonctions et interactions au sein de systèmes vivants. De nombreuses méthodes ont été proposées pour la recherche d'homologie entre séquences biologiques, permettant la classification de séquences dans des classes fonctionnelles ou structurelles. La recherche d'homologie est depuis plus de 20 ans basée sur la recherche d'alignements entre séquences et sur le calcul d'un score d'alignement, via par exemple l'algorithme de Smith-Waterman [SMI 81] ou ses variantes plus rapides BLAST [ALT 90] et FASTA [PEA 90]. A partir des années 1990, de meilleures performances ont été obtenues par des méthodes construisant des modèles pour différentes classes de séquences, et comparant une séquence à classer à ces différents modèles. Ces modèles incluent par exemple la méthode des profiles [GRI 90], les chaînes de Markov cachées [KRO 94, BAL 94], PSI-BLAST [ALT 97] ou SAM-T98 [KAR 98]. Ces méthodes sont dites *génératives*, au sens ou elles créent des modèles pour différentes classes de séquences, et vérifient à quel degré ces modèles expliquent une séquence à classer.

De nouveaux gains en performance ont été réalisés depuis 5 ans, avec l'utilisation de méthodes *discriminantes* pour la classification supervisée. Par opposition aux méthodes génératives, ces méthodes apprennent des règles de classification qui prennent en compte les différences entre classes. Une attention particulière a été portée sur l'uti-

lisation de machines à vecteurs de support (SVM) pour la classification de séquences en familles d'homologues. Les SVM sont des algorithmes d'apprentissage statistique [VAP 98, CRI 00, SCH 02] pour la classification supervisée en différentes classes. Un élément important des SVM est l'utilisation d'une fonction, appelée noyau, pour mesurer la similarité entre n'importe quelle paire d'éléments à classer, des séquences dans notre cas. En utilisant différents noyaux, on peut obtenir une grande variété de SVM avec des performances différentes sur un problème de classification donné. Dans le cas de séquences biologiques, plusieurs noyaux ont été développés au cours des dernières années. La première utilisation des SVM pour la classification de séquences biologiques a été proposée par [JAA 00] à l'aide du noyau de Fisher déduit d'un modèle statistique de séquences. D'autres tentatives incluent la méthode "SVM-pairwise" [LIA 02], ou les noyaux de spectre [LES 02] et de mismatch [LES 03].

Une fonction noyau peut souvent être considérée comme une mesure de similarité entre objets à classer. En particulier, une SVM apprend une fonction telle que des objets "similaires" (au sens de la fonction noyau) tendent à appartenir à des classes similaires. Cette observation suggère que des noyaux intéressants peuvent être construits à partir de mesures de similarité pertinentes. Dans le cas des séquences biologiques, les mesures de similarité par alignements de séquences sont couramment utilisées pour la mesure directe de similarité entre séquences, car elles quantifient de manière naturelle des phénomènes biologiques responsables de l'évolution de séquences (notamment les mutations, insertions, et délétions). Le but de cette contribution est donc d'analyser sous quelles conditions ces mesures de similarité peuvent être utilisées comme fonction noyau par des SVM dans un contexte de classification.

Nous montrons dans un premier temps que les scores d'alignement, même s'il constituent des mesures de similarité intéressantes, ne peuvent pas être utilisés directement par des SVM, car ils ne sont pas définis positifs. Dans un deuxième temps, nous montrons comment des opérations de convolution permettent de construire un noyau défini positif utilisant les mêmes informations que les scores d'alignement. Les preuves des résultats énoncés dans cette contribution et des extensions de ce travail se trouvent dans les références [SAI 03] et [VER 04].

## 2. Machines à vecteurs de support et noyaux positifs

Les SVM pour la classification supervisée sont des algorithmes introduits par Vapnik et ses collègues dans les années 1990 [BOS 92]. Dans le cas de la classification supervisée binaire, une SVM apprend un fonction de classification à partir d'un ensemble d'exemples positifs $\mathcal{X}_+$ et négatifs $\mathcal{X}_-$ de la forme :

$$f(x) = \sum_{i:x_i \in \mathcal{X}_+} \lambda_i K(x, x_i) - \sum_{i:x_i \in \mathcal{X}_-} \lambda_i K(x, x_i), \tag{1}$$

où les poids positifs $\lambda_i$ associés aux exemples d'entraînement sont calculés par maximisation d'une fonctionnelle quadratique. La fonction $K(.,.)$ est appelée un noyau. Une nouvelle séquence $x$ est classée dans la classe positive (resp. négative) si la fonction $f(x)$ est positive (resp. négative). Une présentation plus détaillée de l'algorithme SVM peut être trouvée dans différents ouvrages [VAP 98, CRI 00, SCH 02].

Toute fonction $K(.,.)$ peut être utilisée comme noyau dans (1) à condition d'être symétrique et définie positive, ce qui signifie que pour tout nombre $n$ et tout choix de $n$ séquences $\{x_1, \ldots, x_n\}$, la matrice de taille $n \times n$ définie par $K_{i,j} = K(x_i, x_j)$ doit être symétrique et semi-définie positive. De tels fonctions seront appelées des noyaux de séquences dans la suite. Dans la section suivante, nous montrons comment définir un tel noyau à l'aide du concept d'alignement local, couramment utilisé pour la comparaison de séquences biologiques.

## 3. Alignement local de séquences

Commençons par quelques notations. L'alphabet dans lequel les séquences sont écrites est un ensemble fini $\mathcal{A}$ (de 20 lettres dans le cas de séquences protéiques, ou de 4 lettres dans le cas de séquences nucléiques). Une séquence est une suite finie de lettres, et nous notons $\mathcal{X} = \{\epsilon\} \cup \bigcup_{i=1}^{\infty} \mathcal{A}^i$ l'ensemble des séquences finies sur $\mathcal{A}$, $\epsilon$ représentant la séquence vide. La longueur d'une séquence $x \in \mathcal{X}$ est notée $|x|$, et la concaténation de séquences $x$ et $y$ est notée $xy$.

La notion d'alignement entre séquences, couramment utilisée pour comparer des séquences biologiques, est définie formellement de la manière suivante.

**Définition 1** *Un alignement avec gaps $\pi$ de $p \geq 0$ positions entre deux séquences $x$ et $y$ de $\mathcal{X}$ est une paire de p-uples $\pi = ((\pi_1(1), \ldots, \pi_1(p)), (\pi_2(1), \ldots, \pi_2(p))) \in \mathbb{N}^{2p}$ qui satisfait :*

$$1 \leq \pi_1(1) < \pi_1(2) < \ldots < \pi_1(p) \leq |x|,$$
$$1 \leq \pi_2(1) < \pi_2(2) < \ldots < \pi_2(p) \leq |y|.$$

Une manière courante de représenter un alignement entre deux séquences est de les écrire l'une au-dessus de l'autre, en alignant les lettres définie par l'alignement et en rajoutant des signes '-' pour représenter les gaps. Par exemple, si $x =$ `GAATCCG` et $y =$ `GATTGC`, alors l'alignement de 4 lettres $\pi = ((1, 2, 4, 6), (1, 3, 4, 5))$ est représenté par :

```
G-AATCCG-
GAT-T-G-C
```

Soit $\Pi(x, y)$ l'ensemble des alignements entre deux séquences $x$ et $y$, et soit $|\pi|$ le nombre de lettres alignées dans l'alignement $\pi \in \Pi(x, y)$. Afin de trouver un "bon" alignement entre deux séquences, différentes fonctions de score $s : \Pi(x, y) \to \mathbb{R}$ ont été développées, parmi lesquelles le score d'alignement local défini formellement comme suit :

**Définition 2** *Etant donné une matrice de substitution $S \in \mathbb{R}^{\mathcal{A} \times \mathcal{A}}$ et une fonction de pénalité de gaps $g : \mathbb{N} \to \mathbb{R}$ telle que $g(0) = 0$, on défi nit le score d'alignement local d'un alignement $\pi \in \Pi(x, y)$ par :*

$$s_{S,g}(\pi) := \sum_{i=1}^{|\pi|} S\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right) - \sum_{i=1}^{|\pi|-1} \left[g(\pi_1(i+1) - \pi_1(i) - 1) + g(\pi_2(i+1) - \pi_2(i) - 1)\right]. \quad (2)$$

En d'autres termes, le score d'alignement local de $\pi$ est la somme des scores de substitutions entre lettres alignées, moins la somme des pénalités de gaps quand des gaps sont présents. De ce score, on déduit le score d'alignement entre deux séquences :

**Définition 3** *Le score d'alignement local, ou score de Smith-Waterman (noté score SW) entre deux séquences $(x, y) \in \mathcal{X}^2$ est le score d'alignement local de leur meilleur alignement, i.e.,*

$$SW_{S,g}(x, y) := \max_{\pi \in \Pi(x,y)} s_{S,g}(\pi). \quad (3)$$

Le score de SW est couramment utilisé pour mesurer la similarité entre séquences, et peut être calculé avec une complexité $O(|x||y|)$ par programmation dynamique [SMI 81].

Ce score étant une mesure de similarité "naturelle" entre séquences biologiques, il est naturel de se demander si il peut être utilisé comme noyau par des SVM. Etant clairement symétrique, il suffit de vérifier s'il est défini positif ou non. Des résultats expérimentaux montrent que la réponses est négative en général, en particulier pour des matrices de similarité et des pénalités de gaps utilisées en pratique : il est possible de trouver des ensembles de séquences telles que la matrice de similarité résultante ait des valeurs propres négatives. Comme le montre la proposition suivante, ce résultat négatif dépend cependant des paramètres choisi :

**Proposition 1** *Soit $g = 0$ (pas de pénalité de gap) et $S$ la matrice de substitution nulle sauf pour une lettre $a \in \mathcal{A}$ sur la diagonale, i.e., $S(a, a) = 1$ et $S(u, v) = 0$ sauf si $u = v = a$. Alors le score $SW_{S,g}$ est un noyau pour séquence défi ni positif.*

## 4. Noyau d'alignement local

Afin d'utiliser la notion d'alignement local avec des SVM, nous allons maintenant définir des noyaux définis positifs à partir de scores d'alignement. Notre travail repose sur une opération définie par [HAU 99] laissant invariant l'espace des noyaux définis positifs sur un ensemble : la convolution. Dans le cas de noyaux pour séquences, la convolution est l'opération qui à deux noyaux $K_1$ et $K_2$ associe le noyau pour séquence $K_1 \star K_2$ défini par :

$$\forall (x,y) \in \mathcal{X}^2, \quad K_1 \star K_2(x,y) = \sum_{x_1 x_2 = x, y_1 y_2 = y} K_1(x_1,y_1) K_2(x_2,y_2).$$

Si $K_1$ et $K_2$ sont des noyaux de séquences définis positifs, alors leur convolution $K_1 \star K_2$ est également un noyau défini positif [HAU 99]. Pour tout noyau de séquences $K$, on note $K^{(n)}$ le noyau obtenu par $n$ convolutions de $K$ avec lui-même.

Les noyaux de convolution ainsi définis sont utiles pour comparer des séquences de différentes longueurs, mais qui ont des parties communes. Par exemple, [WAT 00] et [HAU 99] montrent que la probabilité d'émettre deux séquences par une "pair-HMM" est un noyau de convolution, et peut donc être utilisé comme noyau par les SVM. Nous allons à présent étendre cette idée pour définir, par convolution, un noyau qui imitent des mesures de similarité par recherche d'alignement local.

Pour cela, nous allons définir formellement trois noyaux de séquence de base. Le premier est un noyau trivial, toujours égal à 1 :

$$\forall (x,y) \in \mathcal{X}^2, \quad K_0(x,y) = 1.$$

Deuxièmement, afin de quantifier l'alignement entre deux lettres, nous définissons le noyau :

$$K_a^{(\beta)}(x,y) = \begin{cases} 0 & \text{if } |x| \neq 1 \text{ or } |y| \neq 1, \\ \exp\left(\beta S(x,y)\right) & \text{otherwise,} \end{cases} \tag{4}$$

où $\beta \geq 0$ est un paramètre and $S : \mathcal{A}^2 \to \mathbb{R}$ est une matrice de similarité symétrique telle que la matrice $\left(\exp\left(\beta S(a,b)\right)\right)_{a,b \in \mathcal{A}}$ soit semi-définie positive (ce qui est par exemple le cas pour tout $\beta$ si $S$ est conditionnellement définie positive [BER 84]).

Troisièmement, nous définissons le noyau suivant pour quantifier la pénalité des gaps :

$$K_g^{(\beta)}(x,y) = \exp\left[\beta\left(g(|x|) + g(|y|)\right)\right],$$

où $\beta \geq 0$ est un paramètre et $g(n)$ est le coût d'un gap de longueur $n$ donné par :

$$\begin{cases} g(0) = 0 & \text{if } n = 0, \\ g(n) = d + e(n-1) & \text{if } n \geq 1. \end{cases} \tag{5}$$

$d$ et $e$ sont des paramètres appelés coût d'ouverture et d'extension.

Il est facile de vérifier que ces trois noyaux sont bien des noyaux de séquences définis positifs. Il en résulte que le noyau suivant, défini pour tout $n \in \mathbb{N}$ est également défini positif :

$$K_{(n)}^{(\beta)}(x,y) = K_0 \star \left(K_a^{(\beta)} \star K_g^{(\beta)}\right)^{(n-1)} \star K_a^{(\beta)} \star K_0.$$

Ce noyau quantifie la similarité entre deux séquences $x$ et $y$ à travers des alignements de exactement $n$ lettres. En effet, l'opération de convolution consiste à sommer sur toutes les décompositions de $x$ et $y$ en une parties initiales (dont la similarité est mesurée par $K_0$), puis une succession de $n$ lettres (dont la similarité est mesurée par $K_a^{(\beta)}$) éventuellement séparées par $n-1$ gaps (dont la similarité est mesurée par $K_g^{(\beta)}$), puis des parties terminales (dont la similarité est mesurée par $K_0$).

Afin de prendre en compte des alignement d'un nombre quelconque de lettres, nous définissons finalement le noyau suivant, appelé *noyau d'alignement local* :

$$K_{LA}^{(\beta)} = \sum_{i=0}^{\infty} K_{(i)}^{(\beta)}. \tag{6}$$

Ce noyau est défini comme une limite ponctuelle de noyaux définis positifs, et est donc lui-même bien défini positif [BER 84]. L'intérêt de ce noyau réside dans le théorème suivant, qui le relie au score d'alignement local :

**Théorème 1** *Le noyau d'alignement local s'écrit en fonction du score d'alignement local de la manière suivante :*

$$K_{LA}^{(\beta)}(x,y) = \sum_{\pi \in \Pi(x,y)} \exp\left(\beta s_{S,g}(x,y,\pi)\right). \tag{7}$$

*En particulier, le score de SW peut être vu comme une limite quand $\beta$ tend vers l'infi ni :*

$$\lim_{\beta \to +\infty} \frac{1}{\beta} \ln K_{LA}^{(\beta)}(x,y) = SW_{S,g}(x,y). \tag{8}$$

Ces équations clarifient le lien entre le noyau d'alignement local et le score de SW, et mettent en évidence pourquoi ce score n'est pas défini positif. Premièrement, le score de SW ne conserve que la contribution du meilleur alignement, alors que le noyau fait une somme sur tous les alignements. Deuxièmement, le score de SW est le logarithme (à la limite) d'un noyau défini positif, et le passage au logarithme est une opération qui ne conserve pas la propriété d'être défini positif en général [BER 84].

## 5. Implémentation

Une implémentation naïve du noyau d'alignement local à partir de la formule (7) nécessiterait une somme sur $|\Pi(x,y)|$ alignements, et résulterait en une complexité exponentielle en $|x|$ et $|y|$. Cependant, tout comme le score de SW, le calcul peut être factorisé par programmation dynamique pour aboutir à une implémentation en $O(|x||y|)$ (voir détails dans [VER 04]).

Dans la pratique, cependant, ce noyau souffre comme d'autres noyaux pour séquences du problème de la dominance de la diagonale, c'est-à-dire du fait que $K_{LA}^{(\beta)}(x,x)$ peut couramment être des ordres de magnitude plus grand que $K_{LA}^{(\beta)}(x,y)$ pour deux séquences $x$ et $y$. Cela est particulièrement vrai pour les grandes valeurs du paramètre $\beta$, car :

$$\frac{K_{LA}^{(\beta)}(x,x)}{K_{LA}^{(\beta)}(x,y)} \sim \exp\beta\left(SW_{S,g}(x,x) - SW_{S,g}(x,y)\right)$$

quand $\beta \to \infty$. Dans la pratique, il est connu que les SVM ne fournissent pas de bon résultats dans ce cas, car l'apprentissage consiste essentiellement à mémoriser les données observées et la généralisation revient essentiellement à rechercher le plus proche voisin.

Afin d'utiliser le noyau d'alignement local en pratique, nous proposons de prendre son logarithme via la formule suivante :

$$\tilde{K}_{LA}^{(\beta)}(x,y) = \frac{1}{\beta} \ln K_{LA}^{(\beta)}(x,y). \tag{9}$$

Cette opération pose problème, car $\tilde{K}_{LA}^{(\beta)}$ risque de ne pas être défini positif. Dans la pratique, la matrice de similarité entre exemple d'apprentissage utilisée par les SVM risque de posséder des valeurs propres négatives. Pour remédier à ce problème, nous proposons de retrancher à la diagonale de cette matrice la plus petite valeur propre (si elle est négative), afin que la matrice devienne semi-définie positive. Cette astuce n'est bien sûr utile que dans la phase d'apprentissage.

## 6. Expériences et conclusion

Nous avons testé le noyau d'alignement local dans un problème de classification de séquences de domaines protéiques en super-familles de la base de données SCOP [MUR 95] version 1.53. Nous avons suivi l'expérience décrite dans [LIA 02]. Les données[1] consistent en 4352 séquences groupées en familles et super-familles. Pour chaque famille, les séquences de cette familles sont des exemples de test positifs, et les séquences de la même super-famille mais de familles différentes sont les exemples positifs d'entraînement. Les exemples négatifs sont pris en dehors de la super-famille, et sont séparés aléatoirement en exemples d'entraînement et de test. En ne considérant que les familles avec au moins 10 exemples positifs en entraînement et 5 en test, on aboutit à 54 familles. Pour chaque famille, la surface sous la courbe des vrai positifs contre les faux positifs (courbe ROC), normalisée entre 0 et 1, est calculée (indice ROC). De même, la surface sous cette courbe jusqu'à 50 faux positifs est calculée (ROC50), ainsi que le nombre de faux positifs ayant un score supérieur au score médian des vrais positifs (RFP).

Le noyau d'alignement local est comparé avec 3 autres noyaux représentant l'état de l'art en classification supervisée de séquences protéiques : le noyau de Fisher [JAA 00], le noyau "pairwise" [LIA 02], et le noyau mismatch [LES 03].

La table 1 résume les résultats obtenus pour différentes valeurs de $\beta$, ainsi que les scores obtenus par les autres méthodes testées. Ces résultats montrent que les meilleurs résultats sont obtenus quand $\beta$ est de l'ordre de

| Kernel | Mean ROC | Mean ROC50 | Mean mRFP |
|---|---|---|---|
| LA ($\beta = +\infty$) | 0.908 | 0.591 | 0.0654 |
| LA ($\beta = 1$) | 0.912 | 0.612 | 0.0626 |
| LA ($\beta = 0.8$) | 0.908 | 0.597 | 0.0679 |
| LA ($\beta = 0.5$) | **0.925** | **0.649** | 0.0541 |
| LA ($\beta = 0.2$) | 0.923 | **0.661** | 0.0637 |
| LA ($\beta = 0.1$) | 0.868 | 0.429 | 0.111 |
| Pairwise | 0.896 | 0.464 | 0.0837 |
| Mismatch | 0.872 | 0.400 | 0.0837 |
| Fisher | 0.773 | 0.250 | 0.204 |

**TAB. 1.** *ROC, ROC50 et RFP moyens obtenus sur 54 familles pour différentes noyaux. LA-eig représente le noyau d'alignement local. $\beta = +\infty$ correspond au score de SW.*

$0.2 - 0.5$, et qu'ils sont meilleurs que l'état de l'art représenté par les autres noyaux. Les distributions des scores ROC, ROC50 et RFP sur les 54 familles pour différentes noyaux sont montrés sur les figures 1, 2 et 3. Ces résultats illustrent d'une part l'intérêt d'utiliser une mesure de similarité naturelle pour obetnir de bonnes performance en classification, et d'autre part le gain obtenu en prenant en compte l'ensemble des alignements entre deux séquences plutôt que le meilleur alignement uniquement.

## 7. Bibliographie


[ALT 90] ALTSCHUL S., GISH W., MILLER W., MYERS E., LIPMAN D., A basic local alignment search tool, *Journal of Molecular Biology*, vol. 215, 1990, p. 403–410.

[ALT 97] ALTSCHUL S., MADDEN T., SCHAFFER A., ZHANG J., ZHANG Z., MILLER W., LIPMAN D., Gapped BLAST and PSI-BLAST : A new generation of protein database search programs, *Nucleic Acids Research*, vol. 25, 1997, p. 3389–3402.

[BAL 94] BALDI P., CHAUVIN Y., HUNKAPILLER T., MCCLURE M., Hidden Markov models of biological primary sequence information, *Proc. Natl. Acad. Sci. USA*, vol. 91(3), 1994, p. 1053–1063.


1. Accessibles à `www.cs.columbia.edu/compbio/svm-pairwise`

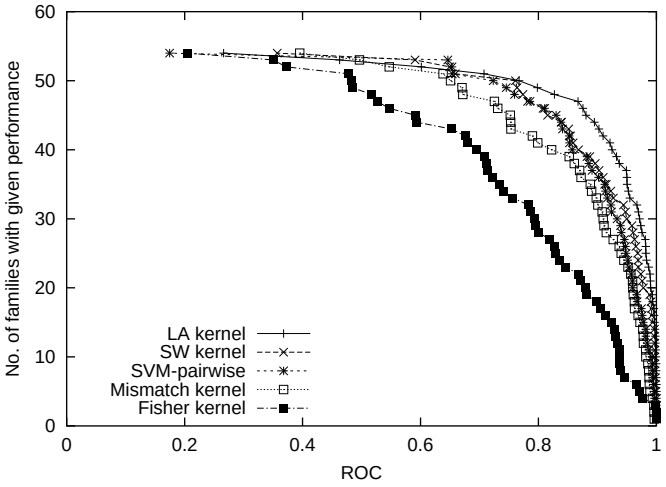


**FIG. 1.** *Distribution du score ROC pour différents noyaux. La courbe noté "LA kernel" correspond au noyau d'alignement local avec $\beta = 0.5$.La courbe "SW kernel" correspond au score de SW.*

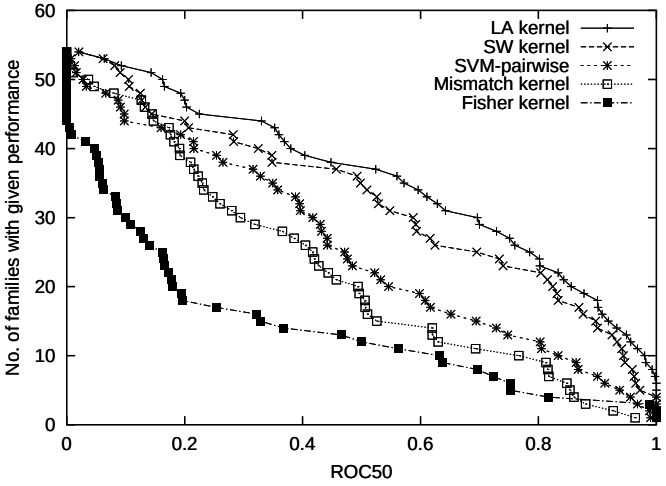


**FIG. 2.** *Distribution du score ROC50 pour différents noyaux.*


[BER 84] BERG C., CHRISTENSEN J., RESSEL P., *Harmonic analysis on semigroups*, Springer-Verlag, New-York, 1984.

[BOS 92] BOSER B. E., GUYON I. M., VAPNIK V. N., A training algorithm for optimal margin classifiers, *Proceedings of the 5th annual ACM workshop on Computational Learning Theory*, ACM Press, 1992, p. 144–152.

[CRI 00] CRISTIANINI N., SHAWE-TAYLOR J., *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[GRI 90] GRIBSKOV M., LÜTHY R., , EISENBERG D., Profile Analysis, *Methods in Enzymology*, vol. 183, 1990, p. 146–159.

[HAU 99] HAUSSLER D., Convolution Kernels on Discrete Structures, rapport, 1999, UC Santa Cruz.

[JAA 00] JAAKKOLA T., DIEKHANS M., HAUSSLER D., A Discriminative Framework for Detecting Remote Protein Homologies, *Journal of Computational Biology*, vol. 7, n$^o$ 1,2, 2000, p. 95–114.

[KAR 98] KARPLUS K., BARRETT C., HUGHEY R., Hidden Markov Models for Detecting Remote Protein Homologies, *Bioinformatics*, vol. 14, n$^o$ 10, 1998, p. 846–856.

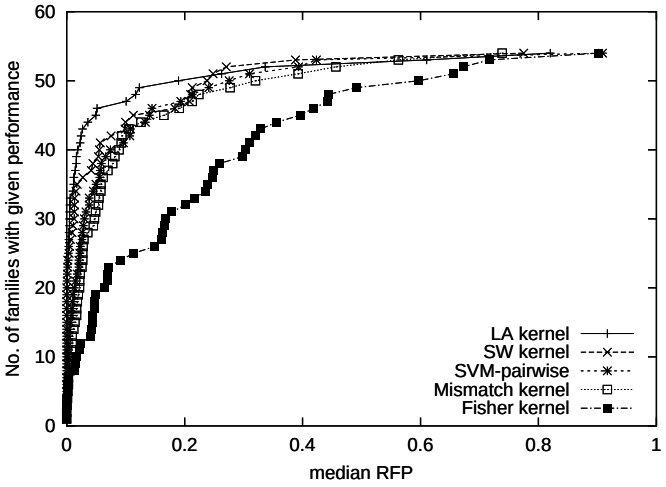


**FIG. 3.** *Distribution du score RFP pour différents noyaux.*


[KRO 94] KROGH A., BROWN M., MIAN I., SJOLANDER K., HAUSSLER D., Hidden Markov models in computational biology : Applications to protein modeling, *Journal of Molecular Biology*, vol. 235, 1994, p. 1501–1531.

[LES 02] LESLIE C., ESKIN E., NOBLE W. S., The spectrum kernel : a string kernel for SVM protein classification, ALTMAN R. B., DUNKER A. K., HUNTER L., LAUERDALE K., KLEIN T. E., Eds., *Proceedings of the Pacific Symposium on Biocomputing 2002*, World Scientific, 2002, p. 564–575.

[LES 03] LESLIE C., ESKIN E., WESTON J., NOBLE W. S., Mismatch String Kernels for SVM Protein Classification, BECKER S., THRUN S., OBERMAYER K., Eds., *Advances in Neural Information Processing Systems 15*, MIT Press, 2003.

[LIA 02] LIAO L., NOBLE W. S., Combining pairwise sequence similarity and support vector machines for remote protein homology detection, *Proceedings of the Sixth International Conference on Computational Molecular Biology*, ACM Press, 2002, p. 225–232.

[MUR 95] MURZIN A., BRENNER S., HUBBARD T., CHOTHIA C., SCOP : A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, vol. 247, 1995, p. 536–540.

[PEA 90] PEARSON W., Rapid and sensitive sequence comparisons with FASTP and FASTA, *Methods in Enzymology*, vol. 183, 1990, p. 63–98.

[SAI 03] SAIGO H., VERT J.-P., UEDA N., AKUTSU T., Protein homology detection using string alignment kernels, *Bioinformatics*, , 2003, To appear.

[SCH 02] SCHÖLKOPF B., SMOLA A. J., *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2002.

[SMI 81] SMITH T., WATERMAN M., Identification of common molecular subsequences., *Journal of Molecular Biology*, vol. 147, 1981, p. 195-197.

[VAP 98] VAPNIK V. N., *Statistical Learning Theory*, Wiley, New-York, 1998.

[VER 04] VERT J.-P., SAIGO H., AKUTSU T., Convolution and local alignment kernels, SCHÖLKOPF B., TSUDA K., VERT J.-P., Eds., *Kernel Methods in Computational Biology*, The MIT Press, 2004, (to appear).

[WAT 00] WATKINS C., Dynamic alignment kernels, SMOLA A., BARTLETT P., SCHÖLKOPF B., SCHUURMANS D., Eds., *Advances in Large Margin Classifiers*, p. 39–50, MIT Press, Cambridge, MA, 2000.