Traitement Statistique des Résultats de SELEX

Damien EVEILLARD[†‡] Yann GUERMEUR[†]

[†] LORIA (UMR 7503) – Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex
[‡] MAEM (UMR 7567) – Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex

Courriel : {Damien.Eveillard,Yann.Guermeur}@loria.fr

Résumé

Les expériences de SELEX constituent une source d'informations particulièrement précieuse en biologie moléculaire. Cependant, les données produites, qui peuvent présenter une variabilité importante, ne sont pas d'une exploitation aisée. Après avoir mis en lumière le fait que l'approche standard de cette exploitation peut conduire à des résultats incomplets, voire erronés, cet article propose de pallier les insuffisances rencontrées en introduisant une étape de traitement supplémentaire, de nature statistique. Cette étape, qui a pour but d'identifier le nombre de motifs à déterminer, ainsi que les ensembles de ligands qui leur sont associés, repose sur l'analyse de données et la classification. Elle précède la détermination des séquences consensus. Notre approche est validée sur un cas réel, pour lequel elle produit des résultats permettant de formuler des hypothèses confirmées a posteriori *par une étude biologique indépendante de la nôtre.*

Mots clés : *Données SELEX, classification, protéines SR, alignements multiples.*

1 Introduction

L'essor actuel de la biologie moléculaire appelle l'utilisation de techniques de pointe permettant de traiter automatiquement les grandes masses de données issues de l'expérimentation. La méthode du SELEX, destinée à mettre en évidence les molécules ligands se fixant par affinité sur une molécule cible donnée, entre dans cette catégorie. Son emploi est actuellement largement répandu. Si les données qu'elle produit sont riches d'information, leur exploitation est délicate. Nous soulignons ici les limites de l'approche standard, consistant à dériver une ou plusieurs séquences consensus à partir de l'ensemble des ligands obtenus. Cette approche peut avoir pour conséquence, suivant le cas, soit de produire des séquences consensus sans signification biologique (ne correspondant pas à un motif), soit de négliger des motifs. Afin de pallier ces insuffisances, nous proposons un prétraitement statistique des données SELEX aboutissant à un partitionnement des ligands. Les consensus sont alors calculés indépendamment pour chaque classe. Cette approche est illustrée sur un cas réel, celui de la protéine SR 9G8. Dans ce cas, les ligands obtenus expérimentalement ne possèdent pas de structures secondaires caractéristiques. Ainsi, nous traitons la situation la plus délicate, dans laquelle aucune connaissance biologique n'est disponible a priori pour orienter la recherche. Des arguments purement statistiques permettent dans ce cas de formuler une conjecture de nature biologique, l'existence de motifs multiples, qui se trouve validée par les observations résultant d'expériences effectuées indépendamment par une autre équipe.

L'organisation de cet article est la suivante. La section 2 contient un bref rappel sur le principe et la mise en œuvre de la technique du SELEX. La section 3 décrit l'exploitation standard des données SELEX et met en évidence ses imperfections. La classification des ligands et son influence sur la détermination des séquences consensus sont étudiées dans la section 4. La section 5 est dédiée à la validation biologique de notre principale conjecture.

2 Principe des expériences de SELEX

La technique expérimentale de *Systematic Evolution of Ligands by EXponential enrichment* (SELEX) [19] est une méthode du domaine de la chimie combinatoire qui permet d'extraire des ligands potentiels d'une banque d'oligonucléotides composée initialement de séquences engendrées aléatoirement. La nature des ligands obtenus résulte conjointement des affinités et des concentrations des molécules cibles. Le SELEX est une méthode itérative constituée de quatre étapes schématisées sur la figure 1.

Initialement, un ensemble d'acides nucléiques est engendré aléatoirement. La mise en présence de ces acides nucléiques avec une molécule cible produit des liaisons biochimiques dont l'intensité est fonction du couple considéré. Parmi les complexes constitués, on sélectionne ceux dont la liaison est suffisamment forte. Les ligands contenus dans les complexes sont ensuite identifiés puis amplifiés par *Polymerase Chain Reaction* (PCR). Un nouvel ensemble d'acides nucléiques est ainsi produit, sur lequel il est possible d'itérer le processus de sélection. L'expérimentateur décide de mettre fin à la succession des cycles lorsqu'il lui apparaît que les données obtenues correspondent aux critères biologiques fixés (forte affinité pour la molécule cible).

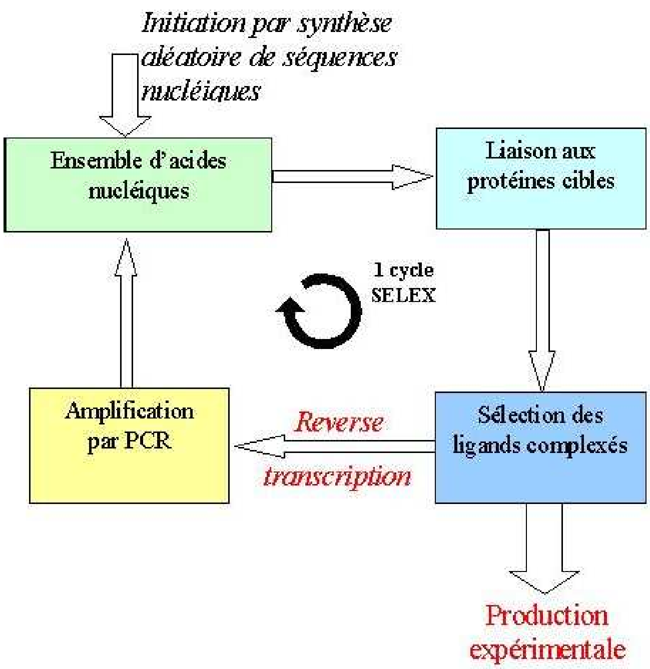

FIG. 1 – Etapes de la méthode SELEX

Dans ce qui suit, dans un but de simplification, l'exposé est restreint au cadre dans lequel se place notre étude, celui des interactions ARN-protéines. La méthode permet alors d'identifier les fragments d'ARN possédant un motif reconnu par la protéine cible. Sous l'hypothèse que ce motif est unique, il sera le représentant biologique du critère de sélection et sera contenu dans chaque séquence issue de l'expérience.
Du fait de son utilité et de sa performance, la méthode SELEX a suscité plusieurs analyses théoriques (voir par exemple [20, 16]). Elles établissent en particulier les faits suivants. La phase d'amplification des oligonucléotides conserve les proportions des différents ligands. Le partitionnement constitue l'étape expérimentale la plus sensible. L'affinité globale des ligands augmente de manière exponentielle avec le nombre de cycles SELEX. Du fait de la nature itérative de la méthode, les grandeurs d'intérêt (concentration et affinité des différents ligands), se comportent comme des processus stochastiques indicés par le numéro de cycle. Il est possible de déterminer, pour un problème donné, un nombre de cycles optimum. Tandis qu'un nombre trop faible de cycles ne permet pas de filtrer suffisamment les oligonucléotides, un nombre trop élevé réduit trop fortement la diversité des ligands obtenus.

3 Limites de l'exploitation standard des données du SELEX

Dans cette section, nous nous appuyons sur un exemple particulier d'expériences SELEX, représentant un enjeu biologique important, pour mettre en évidence les limites de l'approche standard du traitement des données produites.

3.1 Description de l'analyse standard

Chaque ligand issu d'une expérience de SELEX possède une instance d'au moins un motif satisfaisant le critère de sélection. La question se pose de produire un représentant synthétique d'un tel motif. L'approche la plus judicieuse consiste à effectuer un alignement multiple de type local [8]. L'algorithme MM [2, 1], qui repose sur le principe inductif de maximum de vraisemblance, est dans ce cadre expérimental le plus utilisé. Il permet de mettre rapidement en évidence le ou les motifs communs aux séquences biologiques qui lui sont fournies, que ces séquences possèdent ou non la même taille. La représentation synthétique des motifs est constituée par l'ensemble des séquences consensus qui sont retournées.

3.2 Nature des insuffisances rencontrées

Les protéines SR (voir par exemple [11]) jouent un rôle particulièrement important, puisqu'elles permettent une régulation des sites d'épissage constitutif et alternatif des sytèmes biologiques hautement différenciés. Elles sont très conservées en taille et en séquence dans tous les organismes chez lesquels elles existent. Leur rôle ne se limite pas à la réaction d'épissage, puisqu'on les retrouve sur les ARN après maturation. Globalement, elles renforceraient

Protéine SR	Séquences consensus	Auteurs
ASF/SF2	AGGACAGAGC RGAAGAAC	Tacke et Manley [18]
	(G/C)R(G/C)A(G/C)GA	Liu *et al.* [12]
SC35	GUUCGAGUA	Tacke et Manley [18]
	GGGUAUGCUG	Cavaloc *et al.* [4]
	UGCNGYY	Schaal et Maniatis [14]
SRp40	GAGCRGUYRGCUC	Tacke *et al.* [17]
	AC(A/G/U)G(G/C)	Liu *et al.* [12]
9G8	AGAC(G/U)ACGAY ACGAGAGAY	Cavaloc *et al.* [4]
	AGAC(G/U)ACGA(C/U)	Lejeune *et al.* [10]

TAB. 1 – Ensembles de séquences consensus obtenues pour différentes protéines SR. Les parenthèses signalent un choix multiple à une position donnée de la séquence. Les symboles sont standard (R : purine, Y : pyrimidine, N : un des 4 nucléotides).

la stabilité des multiples intéractions ARN/ARN et ARN/protéines au sein de complexes spliceosomaux. Elles possèdent des séquences préférentielles de fixation à l'ARN qui leur permettent de moduler l'utilisation de sites d'épissage alternatif. La recherche des motifs nucléiques est donc au centre de l'étude de leurs fonctions, ce qui en fait de manière duale des candidates idéales pour l'évaluation de l'exploitation des expériences SELEX.
De nombreuses études se sont ainsi appliquées à mettre en évidence les motifs susceptibles de fixer ces protéines. Elles ont pu donner lieu, pour une même protéine, à des résultats entièrement différents, comme on pourra le constater dans le tableau 1.
Il est possible d'avancer deux hypothèses pour expliquer ces variations. D'une part, les consensus peuvent correspondre à autant de motifs (possédant la même structure secondaire). Dans ce cas, le fait que la méthode ne les mette pas systématiquement tous en évidence constitue une faiblesse non négligeable. Un défaut plus important encore apparaît dans le cas où certains consensus ne correspondent pas à des motifs. Cette situation peut survenir lorsque différents motifs sont répartis sur des groupes de séquences différents que l'alignement ne distingue pas. Les consensus se presentent alors comme des résultats d'interférences qui n'ont pas eux-mêmes de sens biologique. Ces deux hypothèses appellent la réalisation d'études supplémentaires en préambule à l'alignement. C'est dans cette optique que nous proposons d'effectuer une analyse statistique des données issues du SELEX.

3.3 Caractérisation statistique des insuffisances

Il découle de ce qui précède que l'ensemble des consensus produits par la méthode doit être un invariant en bijection avec l'ensemble des motifs recherchés. Les résultats ne doivent pas être sensibles aux fluctuations expérimentales, comme les erreurs conduisant à l'élimination d'une ou plusieurs séquences nucléiques. Nous proposons d'évaluer cette propriété sur une des protéines SR du tableau 1, la protéine 9G8.
Nous disposons pour cela de données SELEX d'affinité pour la protéine 9G8 obtenues après 11 cycles de sélection [10]. Dans le cadre de notre étude, cette protéine possède l'"avantage" de ne pas fixer de structure secondaire particulière. Elle apparaît donc comme une candidate idéale pour une analyse reposant uniquement sur l'information de séquence. Les 56 séquences nucléiques produites possèdent une longueur variant entre 18 et 22 bases nucléiques. L'application de MM sur l'ensemble de ces données produit un seul consensus significatif au sens du critère utilisé (issu de la théorie de l'information), `AGAC(U/A)ACG`. Ce consensus, que nous qualifierons par la suite de total, est pratiquement identique au consensus commun aux études de Cavaloc et Lejeune (voir le tableau 1). Pour tester la robustesse de la méthode de SELEX, on s'intéresse à la variabilité du consensus obtenu lorsque la base de données est amputée d'une à deux séquences. Les sous-ensemble de la base à 55 et 54 élements étant trop nombreux pour être tous considérés, on extrait de cet ensemble un sous-ensemble représentatif de petite taille. Le critère utilisé pour caractériser la représentativité est la similarité de la ou des séquences consensus produites à la séquence consensus totale. Cette mesure est donnée par le score renvoyé par l'algorithme d'alignement global utilisé ici, celui de Needleman et Wunsch [13]. Plus précisément, on cherche à rendre compte du spectre des similarités obtenues, en assurant en particulier la présence de représentants parmi les sous-ensembles correspondant aux consensus divergeant le plus fortement. La sélection résulte d'un tirage aléatoire sans remise selon une distribution uniforme. Un raisonnement simple permet d'établir qu'en tirant ainsi 59 sous-bases, la probabilité qu'aucun consensus résultant ne se trouve parmi les 5% les plus divergeants est inférieure à 5%. Ce protocole expérimental

est résumé sur la figure 2.

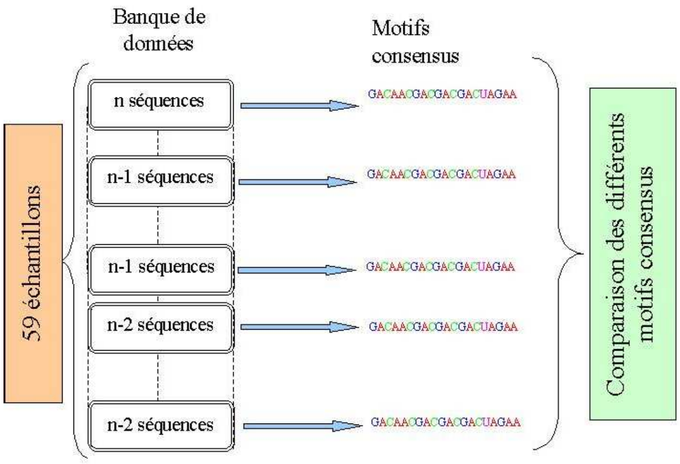

FIG. 2 – Protocole d'étude de la variabilité du consensus associé aux ligands issus d'une expérience de SELEX

Une sélection des consensus "partiels" obtenus est présentée dans le tableau 2. Elle est caractéristique des différents niveaux de similarité entre ces consensus et le consensus total.

Consensus pour la protéine 9G8
AGAC(U/A)ACG (total)
...
ACGACGAU
...
ACGACGAU
...
CUACGCUA
GACGAGAGAUU

TAB. 2 – Consensus total et consensus partiels obtenus après rééchantillonnage. Ces derniers sont classés par ordre de similitude décroissante avec le consensus total.

Les scores de l'alignement des deux derniers consensus partiels avec le consensus total sont respectivement égaux à -4 et -6. L'analyse biologique, réalisée par Fabrice Leclerc, établit que ces deux consensus se distinguent significativement des autres. La présence de cette paire constitue un contre exemple démontrant que les résultats d'analyse SELEX pour la protéine 9G8 ne sont pas consistants. L'algorithme peut produire des consensus très différents pour des conditions expérimentales très proches. Ainsi que nous l'avons indiqué, une cause de cette instabilité peut être la présence de plusieurs motifs au sein des données SELEX, motif associés à des familles de séquences nucléiques différentes. Un moyen de prendre en compte cette possibilité consiste à rechercher les ensembles caractérisés par un motif, en effectuant une classification.

4 Classification sur les séquences produites par la méthode SELEX

4.1 Choix des algorithmes

La classification de séquences fondée sur les scores d'alignement a souvent permis d'obtenir de bonnes performances en génomique. On pourra par exemple consulter sur le sujet [6, 5]. Nous reprenons ici cette stratégie. Deux méthodes standard de classification sont mises en œuvre sur les consensus, la méthode des k-means (voir entre autres [7]) et l'Analyse en Composantes Principales (ACP) hiérarchique, décrite dans [9]. Les scores issus de l'algorithme de Needleman et Wunsch sont de nouveau employés comme mesure de similarité (la "dissimilarité" entre deux séquence étant égale à l'inverse du score de la programmation dynamique). Notons cependant que des expériences supplémentaires, effectuées avec l'algorithme d'alignement local de Smith et Waterman [15], ont conduit aux mêmes résultats qualitatifs (même structure classificatoire).

4.2 Résultats de la classification

Tandis que la méthode des k-means impose de choisir *a priori* le nombre de catégories, l'ACP hiérarchique produit automatiquement une hiérarchie à partir de laquelle on peut construire une structure classificatoire plus ou moins fine, en fonction du seuil retenu. On observe ici une forte concordance des résultats des deux méthodes, ceci pour différents choix du nombre de clusters. Nous illustrons le phénomène dans le cas le plus pertinent, celui de deux classes. La figure 3 localise ainsi les clusters produits par les k-means sur la hiérarchie issu de l'ACP.

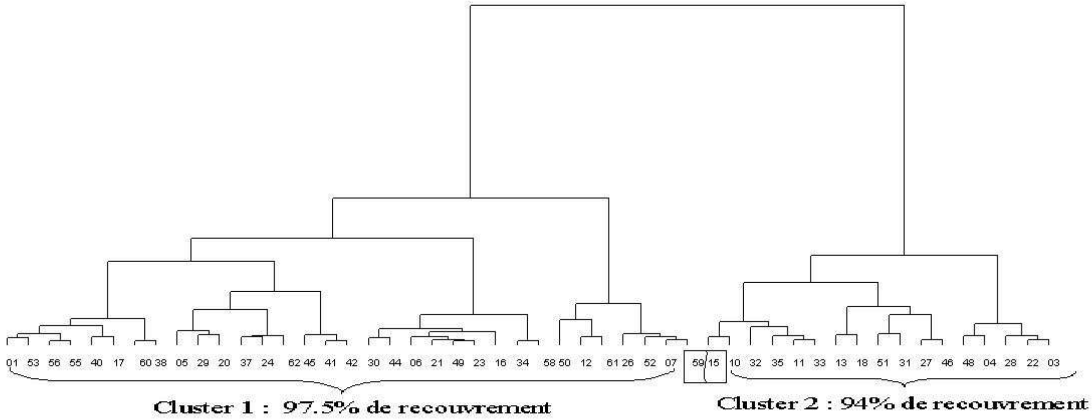

FIG. 3 – Superposition des résultats des deux classifications

Les deux classifications, dont le taux de recouvrement est de 96,4%, répartissent les données entre les deux clusters dans un rapport 3/4 - 1/4. (voir également le tableau 3 contenant la “matrice de confusion” correspondante).

	cluster 1 (k-means)	cluster 2 (k-means)
cluster 1 (ACP)	40	1
cluster 2 (ACP)	1	14

TAB. 3 – Matrice de répartition des séquences entre les clusters issus des deux classifications

La figure 4 propose une représentation alternative, dans laquelle les clusters produits par les k-means sont cette fois matérialisés dans le premier plan principal de l'ACP.

4.3 Analyse des consensus issus de la classification

L'analyse de cette section utilise la classification en deux catégories produite par la méthode des k-means. Les résultats exposés ne varient cependant pas lorsque l'autre classification est retenue. Les clusters identifiés, MM est à nouveau utilisé pour déterminer les consensus correspondants. On retrouve ainsi, avec la séquence `AGAC(U/A)ACG` le motif déjà identifié avant classification. L'autre séquence, `GAC(G/U)A(C/G)(G/A)A`, correspond au second motif recensée dans la littérature [4]. La stabilité des résultats est évaluée suivant le protocole décrit dans la section 3. Une sélection des consensus obtenus, 59 par cluster, apparaît dans le tableau 4.

Cluster 1	Cluster 2
AGAC(U/A)ACG (total)	**GAC(G/U)A(C/G)(G/A)A** (total)
...	...
AGACUACGC	GACGA(G/C)AGAU
...	...
UGACACCG	GACAAGAU
UACGACG	GAC(G/C)AGAG

TAB. 4 – Illustration des séquences consensus obtenues après classification

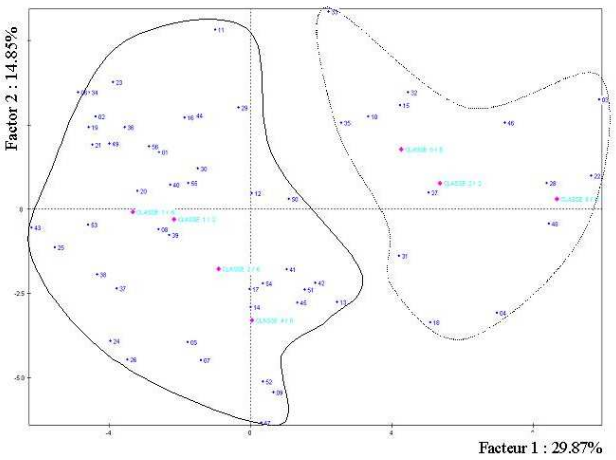

FIG. 4 – Représentation de la classification par k-means dans le premier plan principal de l'ACP

La similitude minimale entre le consensus total et l'un quelconque des consensus partiels est à présent de 1, ceci pour chacun des clusters. Cette valeur est nettement plus élevée que celle obtenue sans classification, -6 (voir la section 3.3). Du fait que les séquences consensus ont toutes approximativement la même taille, la comparaison de ces scores est valide et permet donc de déduire que les données contenues dans les clusters sont plus homogènes. Ici encore, l'expertise biologique vient confirmer les observations numériques, dans la mesure où les séquences issues d'un même cluster apparaissent fonctionnellement similaires. Ceci nous conduit à formuler l'hypothèse que la structure classificatoire déterminée possède une signification biologique et que les deux consensus totaux obtenus sont en fait les représentants de deux motifs différents. Sous cette hypothèse, en analysant les données SELEX suivant la procédure standard, on court le risque de négliger le motif le moins représenté (associé au cluster 2 de notre étude). Ce motif est en quelque sorte masqué par l'autre lorsque l'alignement multiple est effectué.

5 Validation des hypothèses effectuées

La conjecture exprimée dans la section précedente, la présence d'au moins deux motifs, est originale dans le domaine de la biologie fonctionnelle. Jusqu'alors, l'obtention de séquences consensus différentes était seulement imputée aux fluctuations expérimentales. L'hypothèse d'existence de motifs multiples sur des séquences différentes n'était pas retenue. Récemment, l'étude décrite dans [10] a permis d'établir que la protéine 9G8 possède deux domaines. Il s'agit de domaines N-terminaux permettant la fixation aux molécules d'ARN. Ces molécules d'ARN possèdent donc au moins deux types de motifs différents, les interactions multiples permettant à la protéine de moduler l'épissage alternatif et constitutif. Cette observation de nature structurale valide donc qualitativement notre conjecture. Celle-ci est encore corroborée par des observations de nature fonctionnelle. En effet, parmi les protéines SR, la protéine 9G8 est de celles qui contribuent à la modulation de l'épissage tissu-spécifique. Cette fonction multimodale suppose de même la présence de plusieurs motifs. De manière plus générale, la diversité des motifs nous apparaît comme une caractéristique biologique d'une importance centrale. La combinaison SELEX / classification se présente alors comme une solution naturelle permettant de préserver toute l'information biologique pertinente.

6 Conclusions et perspectives

Nous avons exposé une étude critique de la méthode usuelle d'exploitation des données issues de la technique expérimentale du SELEX. Cette méthode, qui soumet sans distinction les données à un alignement multiple, ne permet pas de prendre en compte certains facteurs biologiques importantes, comme la possibilité de l'existence de motifs multiples qui ne seraient pas communs à toutes les séquences. Pour résoudre ce problème, nous proposons d'effectuer sur les séquences produites une classification, et de déterminer ensuite les séquences consensus pour chaque cluster. Dans le cas de la protéine 9G8, cette stratégie nous a permis de mettre en évidence deux motifs différents, ce qui a été confirmé par une étude biologique conduite indépendamment par une autre équipe.
Ces résultats, pour encourageants qu'ils soient, appellent une confirmation qui doit venir d'expériences supplémentaires. Il semble ainsi nécessaire d'appliquer notre approche pour différents nombres de cycles, afin d'étudier l'influence de ce facteur sur le nombre de motifs identifiés. Cette étude est en cours sur un ensemble de protéines SR. La question du choix de la ou des méthodes de classification les plus adaptées est encore pour l'instant lar-

gement ouverte. Un paramètre important doit ici être pris en compte, l'incorporation d'informations biologiques supplémentaires. Ainsi, dans le cas qui nous intéresse plus particulièrement, les relations ARN-protéines peuvent se caractériser par des contraintes très fortes concernant la structure secondaire des ARN. Prendre en compte ces contraintes devrait grandement faciliter la classification. Là encore, des expériences sont en cours. Elles s'appuient sur l'emploi de techniques adaptées aux descripteurs de types différents. La solution que nous privilégions actuellement est le développement de machines à noyaux dédiées, pour mettre en œuvre des algorithmes tels que le *Support Vector Clustering* (SVC) [3].

Remerciements
Les auteurs souhaitent remercier James Stevenin pour la mise à disposition des données expérimentales. Nous exprimons également notre reconnaissance à Fabrice Leclerc pour son analyse des séquences consensus.

Références

[1] BAILEY (T.) et ELKAN (C.), « Fitting a mixture model by expectation maximization to discover motifs in biopolymer », dans PRESS (A.), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994, p. 28–36.

[2] BAILEY (T.) et GRIBSKOV (M.), « Methods and statistics for combining motif match scores », *Journal of Computational Biology*, *5*, 1998, p. 211–221.

[3] BEN-HUR (A.), HORN (D.), SIEGELMANN (H.) et VAPNIK (V.), « Support vector clustering », *Journal of Machine Learning Research*, *2*, 2001, p. 125–137.

[4] CAVALOC (Y.), BOURGEOIS (C. F.), KISTER (L.) et STEVENIN (J.), « The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers », *RNA*, *5*, 1999, p. 468–483.

[5] CLÉMENT (E.) et CODANI (J.-J.), « Lassap, a large scale sequence comparison package », *CABIOS*, *2*, 1997, p. 137–147.

[6] FENG (D. F.) et DOOLITLE (R. F.), « Progressive sequence alignment as a prerequisite to correct phylogenetic trees », *Journal of Molecular Evolution*, *25*, 1987, p. 351–360.

[7] HARTIGAN (J.), *Clustering Algorithms*, Wiley, N.Y., 1975.

[8] L.DURET et ABDEDDAIM (S.), « Multiple alignements for structural, functional, or phylogenetic analyses of homologous sequences », dans HIGGINS (D.) et TAYLOR (W.), *Bioinformatics : Sequence, structure and databanks*, Oxford Univ. Press, 2000, p. 51–76.

[9] LEGENDRE (P.) et LEGENDRE (L.), *Numerical Ecology, second english edition*, Elsevier, 1998.

[10] LEJEUNE (F.), CAVALOC (Y.) et STEVENIN (J.), « Alternative splicing of intron 3 of the serine/arginine-rich protein 9g8 », *Gene*, *276*, nº 11, 2001, p. 7850–7858.

[11] LEWIN (B.), *Gene VII*, Oxford University Press, 2000.

[12] LIU (H.-X.), ZHANG (M.) et KRAINER (A. R.), « Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins », *genes & development*, *12*, 1998, p. 1998–2012.

[13] NEEDLEMAN (S.) et WUNSCH (C.), « A general method applicable to the search for similarities in the amino acid sequence of two proteins », *Journal of Molecular Biology*, *48*, 1970, p. 443–453.

[14] SCHALL (T.) et MANIATIS (T.), « Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA », *Molecular and Cellular Biology*, *19*, 1999, p. 261–273.

[15] SMITH (T.) et WATERMAN (M.), « Identification of common molecular subsequences », *Journal of Molecular Biology*, *47*, nº 1, 1981, p. 195–197.

[16] SUN (F.), GALAS (D.) et WATERMAN (M.), « A mathematical analysis of in vitro molecular selection-amplification », *Journal of Molecular Biology*, *258*, 1996, p. 650–660.

[17] TACKE (R.), CHEN (Y.) et MANLEY (J.), « Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation : Creation of an SRp40-specific splicing enhancer », *Proc. Natl. Acad. Sci. USA*, *94*, 1997, p. 1148–1153.

[18] TACKE (R.) et MANLEY (J. L.), « The Human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities », *EMBO*, *14*, 1995, p. 3540–3551.

[19] TUERK (C.) et GOLD (L.), « Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase », *Science*, *249*, 1990, p. 505–510.

[20] VANT-HULL (B.), PAYANO-BAEZ (A.), DAVIS (R.) et GOLD (L.), « The mathematics of SELEX against complex targets », *Journal of Molecular Biology*, *278*, 1998, p. 579–597.