KOALAB: A new method for regulatory motif search. Illustration on alternative splicing regulation in HIV-1

Damien Eveillard, Abdelhalim Larhlimi, Delphine Ropers, Stéphanie Billaut, Sandrine Peyrefitte

KOALAB: A new method for regulatory motif search. Illustration on alternative splicing regulation in HIV-1

Damien Eveillard[⋆‡], Abdelhalim Larhlimi[⋆], Delphine Ropers[◦] ,
Stéphanie Billaut[⋆], Sandrine Peyrefitte[⋆‡]

[⋆] LORIA, Université Henri Poincaré, BP 239, 54506 Vandœuvre-lès-Nancy, France

[‡] Laboratoire de Maturation des ARN et Enzymologie Moléculaire, UMR 7567 CNRS-UHP, BP 239, 54506 Vandœuvre-lès-Nancy, France

[◦] INRIA-Rhone-Alpes, 655 avenue de l'Europe, 38334 Montbonnot, France

Abstract

Discovering heterogeneous regulatory motifs remains a difficult problem in biological sequence analysis. In this context, statistical learning or pattern search techniques on their own have shown some limitations. However, significant benefits can be taken from their complementarity. We selected two state-of-the-art methods: a multi-class support vector machine (M-SVM) from the statistical learning domain associated with a performant discrete pattern matching algorithm `grappe`, and integrated them into a web technology based graphical software: KOALAB (KOupled Algorithmic and Learning Approach for Biology)[1]. We applied our method on motif discovery within nucleic acid sequences using experimental SELEX results as training database for the M-SVM. An application dealing with the search for splicing regulatory protein binding sites in HIV-1 genome shows the potential of such an approach.

Keywords: motif discovery, multi-class SVM, SELEX, alternative splicing regulation

1 Introduction

Motif discovery within biological sequences is a key area of bioinformatics. Several biological questions have to deal with short sequences mediating interactions between macromolecules such as nucleic acids and/or proteins. Those interactions have been shown to play crucial roles in numerous phenomena. Because of their rather small size, they are difficult to deal with using standard alignment or `Blast` approaches. Pattern search algorithms are remarkably efficient for conserved, fixed motifs. Unfortunately, a growing number of motifs a biologist has to look for do not correspond to that requirement, notably because of intrinsic heterogeneity of the sequences or technical constraints he has to deal with. Indeed, experimental results for biological motif definition will often come as a collection of potential motifs. A common strategy is hence to apply different alignments to extract a global consensus motif from the collection. From a theoretical point

[1] KOALAB 1.0 is freely available at `http://www.loria.fr/equipes/modbio/KOALAB.html`

of view, very often, this type of approach is not satisfactory because of the heterogeneity of the collection so that the deduced consensus does not bear biological reality anymore. Improvements in the efficiency of the methods available and/or alternative computational methods are hence needed to perform the task.

An interesting strategy to define nucleic acid-protein interaction motifs comes from combinatorial chemistry and is named SELEX for Systematic Evolution of Ligands by Exponential Enrichment [21]. The process consists in the selection, among nucleic acid sequences generated randomly, of sequences that bind by affinity to a protein ligand. The result of a SELEX experiment is a collection of less than 100 sequences from which it is, at best, difficult to retrieve any global consensus. In order to take into account the sequence heterogeneity among such a motif collection of biological significance, Roulet et al [19] proposed to use the entire collection as a training database to define Hidden Markov Models (HMM) parameters. A limiting step of this method could be that the model-building process requires sequence alignments as an intermediate step, which can be problematic with small sequences. From a theoretical point of view it means that, before starting the learning process, one has to perform some prior treatment on the data (the alignment) and this is typically not always satisfactory, notably when no biological knowledge can be used to improve the information from the collection.

We propose here to make the most out of a SELEX collection respecting their heterogeneity by using a statistical learning tool: Multiclass Support Vector Machines (M-SVM). These machine learning tools are very useful for discriminant analyses. The multiclass extension of the original binary classifier offers the opportunity, in our context, to take into account, in a single experiment, data corresponding to several proteins. Even if showing some limitations, global consensus interaction sequences have already been independently defined for numerous biological questions, it hence appears interesting to be able to confront them with the results of our method. Furthermore, some other interaction motifs are very conserved and show no need to use statistical methods. In this case, an algorithmic pattern search is the most suitable.

Statistical learning as well as algorithmic methods appear as different techniques that will be more appropriate for different problems, depending on the heterogeneity of the researched signal. In order to be able to deal with any situation and in the way of taking the best benefits from the complementarity of both methods, we propose in this paper an integrated tool that combines them. In order to make it available to the community as a user friendly tool, we thus designed a web technology based graphical software: KOALAB (KOupled Algorithmic and Learning Approach for Biology).

The goal of the present paper is to introduce a new integrated tool for motif research, KOALAB, and to validate it on the complex problem of alternative splicing regulation in HIV-1. The organization is as follows : we briefly describe our integrated approach with the statistical and algorithmic components present in our software (Sect.2). One can submit a consensus motif and/or a motif collection and train the M-SVM, respectively, and retrieve in the same graphical interface the results of both approaches. We illustrate our method (Sect.3) on the search for alternative splicing regulatory sites in HIV-1 using SELEX results for two SR proteins SC35 and 9G8, components of the spliceosome, that will bind such regulatory sites. Our results, confirmed by independant experimental results show the strength of such an integrated approach.

2 KOALAB

2.1 Statistical learning and algorithmic approaches

In order to discover heterogeneous motifs, statistical learning is an efficient approach because it is specified to learn a generalization rule from such type of data. In [20], this method has been used to discover biological motifs such as splicing sites using pattern recognition support vector machine (SVM) [22]. Note, however, that such an approach typically requires several binary classifications for different kinds of motifs. In KOALAB, we propose to use a multi-class support vector machine (M-SVM) [9] to discover several binding motifs simultaneously. In contrast to HMM based methods, we do not need neither a priori knowledge nor prior treatment of the data. The SVM method is a canonical machine learning tool [10, 5] that was proven as a powerful method for pattern recognition in different problems such as handwritten digit recognition, object recognition, voice identification, and text categorization, etc (see for instance [6]). In such areas, the performance of the SVM was equivalent or higher than that of classical non-linear regression models such as neural networks (multilayer perceptrons, MLP). In biology, SVMs have been applied in different fields among which DNA microarrays gene expression data classification [18], protein function classification [2], help on breast cancer diagnosis [14], identification of splicing sites in eukaryotic sequences [20].

Building upon the uniform strong law of large numbers, a new family of multi-class SVMs (M-SVMs) has been specifed in [9]. This specificity allows the scientist to perform a multi-class classification in a single step instead of applying any decomposition method to several binary classification results. In biology, this is of particular interest in order to avoid a too high decomposition level of a complex system.

Conversely, if one has to look for motifs that are more homogeneous i.e., conserved and sufficiently documented, there is no need to use statistical learning methods. This type of motifs may be identified by a discrete pattern matching approach. Hence, we include in our tool an efficient algorithm for pattern search: grappe [13]. Compared to other discrete pattern matching software, grappe offers additional flexibility in pattern description, such as presence of don't care symbols (wildcards) of unbounded or bounded length. Patterns with substitution errors can also be taken into account. The number of errors and their occurrences in a pattern can be specified by the user. A version of `grappe` devoted to nucleic acids sequences treatment (able to deal with specific substitution codes) is used in our tool.

2.2 Integrated software

Although being useful for biological problems, combining the M-SVM and grappe methods remains difficult for a non-specialist. We overcome this problem by developing a software named KOALAB (KOupled Algorithmic and Learning Approach for Biology), which provides a user-friendly interface to M-SVM technology and grappe. KOALAB has been designed to guide the biologist in the discovery of biological motifs in a genome. Using web technology, the user can apply the latest version of the M-SVM software without having to be concerned about technical details. KOALAB can be installed on a local or remote web server. Only a web browser is needed to use it.

The M-SVM interface is designed to reflect the three stages of the process.

- First, the user is invited to provide the learning database containing the collection of motifs together with their respective tags. The learning process can be started.

- Second, the user can perform an evaluation for the progression of the learning process, providing the validation database which is different from the learning one. The evaluation is made on the ability the SVM shows to retrieve the correct labels for these new objects. Even if the theory of such a tool assumes the learning process has to be carried out thoroughly, one can stop it before the strict optimality criteria are satisfied.

- The user can finally proceed with the exploitation phase for which he has to provide the sequence in which to search for motifs. KOALAB provides a practical graphical interface to handle the M-SVM output. The interface incorporates a variety of tools, including data analysis techniques such as signal filtering. Combining them with an M-SVM is an efficient way to detect both known and unknown biological motifs.

When known biological motifs are available, the users can confront the two methods by searching via grappe a pattern corresponding to the motif of interest. KOALAB integrates the statistical and pattern finding results in a graphical representation and summarizes both of them, highlighting the positions along the sequence for which both results are congruent. Because of the high CPU cost of the M-SVM process and given the type of biological questions this tool has been designed for, the users are not supposed to submit entire genomes but rather regions of interest. By combining an algorithmic and a statistical learning method, KOALAB is particularly designed to explore genomes from a new point of view, considering rather small regions in which interactions with regulatory elements lie.

3 Application to regulatory binding sites discovery

As stated before, discovering nucleic acid/protein interaction motifs remains an open problem. We applied our tool to the search for alternative splicing regulatory binding sites in the genome of HIV-1. This phenomenon is crucial for the virus to generate the full protein repertoire during its life cycle (for review, see [11]). As the virus uses the host cell's spliceosomal machinery to allow the splicing of its primary transcript, the splicing regulation mechanisms are similar to that of the cellular pre-messenger RNAs. Among the 9kb long genome sequence, four donor and eight acceptor splice sites allow the virus to produce about 40 different messenger RNAs from which the whole protein repertoire will be obtained. The use of the different sites, i.e. the alternative splicing, is highly regulated in order to produce the right protein at the right time. The regulatory proteins Nef, Tat and Rev are produced in the early phase whereas auxiliary Vip, Vpu, Vpr and structural-enzymatic Gag, Gag-Pol, Env proteins are specific of the late phase. A sharp control of the alternative splicing will drive the progression in the life cycle of the virus. SR (Serine aRginine rich) cellular proteins take part in the splicing reaction [8] consisting in removing introns from the pre-messenger RNA to release the mature messenger RNA. They allow the spliceosome to assemble around a splicing site close to the protein binding sites, this step being often limiting for the whole splicing reaction. Playing such a central part in the process, they are involved in constitutive as well as alternative splicing and have already shown their crucial connection with the HIV-1 virus physiology. We hence searched for SR proteins binding sites KOALAB to this task.

3.1 Training data: ligands from SELEX experiments

The experimental technique of Systematic Evolution of Ligands by EXponential enrichment (SELEX) [21] is a method from combinatorial chemistry devoted to the identification of ligand molecules that bind by affinity on a given target molecule. This experimental approach generates a large number of potential ligands from a nucleic acid pool. The automatically extracted data are safe and without *a priori* assumption, but their interpretation remains difficult especially for the case of heterogeneous data. As stated before, the standard approach, consisting in deriving a consensus motif using different types of alignment methods is often inappropriately applied because of too high heterogeneity in the results. This approach has two major drawbacks: on the one hand, it is possible to produce consensus sequences that have no biological significance (do not correspond to any motif). On the other hand, some significant motifs may be missed. In [19], SELEX experiments are used as a training set to set HMM parameters. In our case, the knowledge on regulatory motifs is not sufficient to develop any probabilistic model which needs some initial conditions that are not always possible to determine. Therefore, a statistical learning approach seems appropriate. To train KOALAB, we used SELEX results for two different SR proteins: SC35 and 9G8. They belong to different subfamilies of SR proteins. SC35 only contains an RRM domain (for RNA Recognition Motif) in its RNA interaction domain whereas 9G8 contains an RRM and a Zn-finger (for Zinc-Finger) motifs [8]. Those data allow to assume that the RNA binding properties of the two proteins will be slightly different and that they won't bind to the same sequences. Prior studies by global consensus determination have confirmed this statement. However, some functional studies suggest that these two proteins might compete one another directly via a competition for some sequence target. That point was particularly interesting to check if our method was able to give some evidence for this situation. The training data were obtained from SELEX experiments with 11 cycles for SC35 and 12 cycles for 9G8 respectively. For each protein, a collection of about 60 sequences (18 nucleotides-long each) was used for training. Compared to the global consensus sequences obtained by other studies for SR proteins, the size of 18nt correspond to the largest motif available, the average being around 10-12 nucleotides. We hence ensure that our windows would encompass the whole interaction region.

3.2 Results

We submitted the whole HIV-1 genome (isolate BRU, genebank accession K02013) to the analysis (it is rather small, about 9kb). We here concentrate on regions of interest more particularly studied because of their importance in the virus alternative splicing regulation. Several splicing regulatory sequences have been identified on the viral RNA downstream of the acceptor A2, A3, A6, A7 sites and upstream from the donor D4 site. We illustrate our results for A3 and A7 regions which are located in structured parts of the viral RNA. The A3 site is exclusively used to produce the tat mRNA and A7 required for the tat and rev mRNA that will give proteins used in the early phase. They are intensively studied because the target of very complex and sharp regulations. As a general comment, we encountered some difficulties with consensus search by `grappe`. Indeed, the consensus available that have been published (for examples [4, 15]) do not even match experimentally proved regions such as the site between ESS3a and ESS3b in the A7 region [17]. The reasons for such a result can be that the consensus are generally longer than the real sequences bound by the proteins or that, as mentioned before, derived

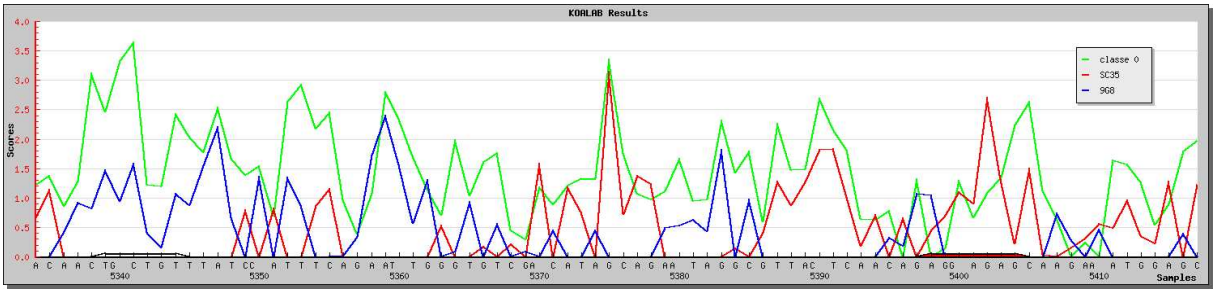

Figure 1: **Areas of alternative splicing regulatory motifs in a part of the HIV-1 genome.** The plots represent the M-SVM results for 2 SR proteins (blue and red) and the non binding segments (green). The bars represent the `grappe` results corresponding to documented SR proteins binding site consensus. The color bar switch to the same SR protein color code when the results for the two methods are consistent.

from a collection of too heterogeneous motifs, the consensus do not correspond to any real sequence anymore. In order to be able to retrieve some matches, one can use central sub-sequences from consensus but soon, the opposite backside effect is observed with the explosion of the system providing a huge quantity of matches. It is anyway of interest to be able to do this confrontation because consensus research was the first technique applied in such type of questions.

An illustration of correlation between grappe and M-SVM result is given in Fig. 1. The graphical interface has been designed to show the grappe results along the sequence as a bar the colour of which corresponds to the M-SVM plots when both results are congruent. In this example covering a region around the A3 site, a SC35 consensus as well as an M-SVM signal for this protein are retrieved around position 5400) where SC35 has effectively been shown to bind (unpublished data). Another consensus grappe hit is found around position 5340 with no M-SVM signal in a region where SC35 is not supposed to bind. Because of the global difficulty to interpret grappe results for SR proteins, we will concentrate here on the presentation of M-SVM results and their confrontation to facts already available on alternative splicing regulation for A3 and A7.

3.2.1 M-SVM results for A3

The A3 site is part from the A3 to A5 sites that are mutually exclusive. A3 is used to produce the tat mRNA whereas A4a, b and c are used for rev mRNA and A5 will take part into env and nef mRNA. We found some sites for SC35 and 9G8 in this region (see Fig. 3.2.2), among which some are experimentally confirmed by a study on SC35 interaction with the viral RNA (unpublished data). An interesting observation is that we find a 9G8 site in a region were SC35 has been proved to bind. This could lead to a competition situation between SC35 and 9G8 that are not equivalent for their effects on the splicing efficiency at the A3 site.

The A3 site activity is also under the control of two Exonic Splicing Silencers, ESS2p and ESS2 that respectively bind hnRNP H and hnRNP A/B repressing proteins. A potential competition between SC35 and hnRNP A/B at the ESS2 site has already been suggested as SC35 has experimentally been shown to bind to ESS2. Our results suggest a potential equivalent situation at the ESS2p site between hnRNP H and 9G8.

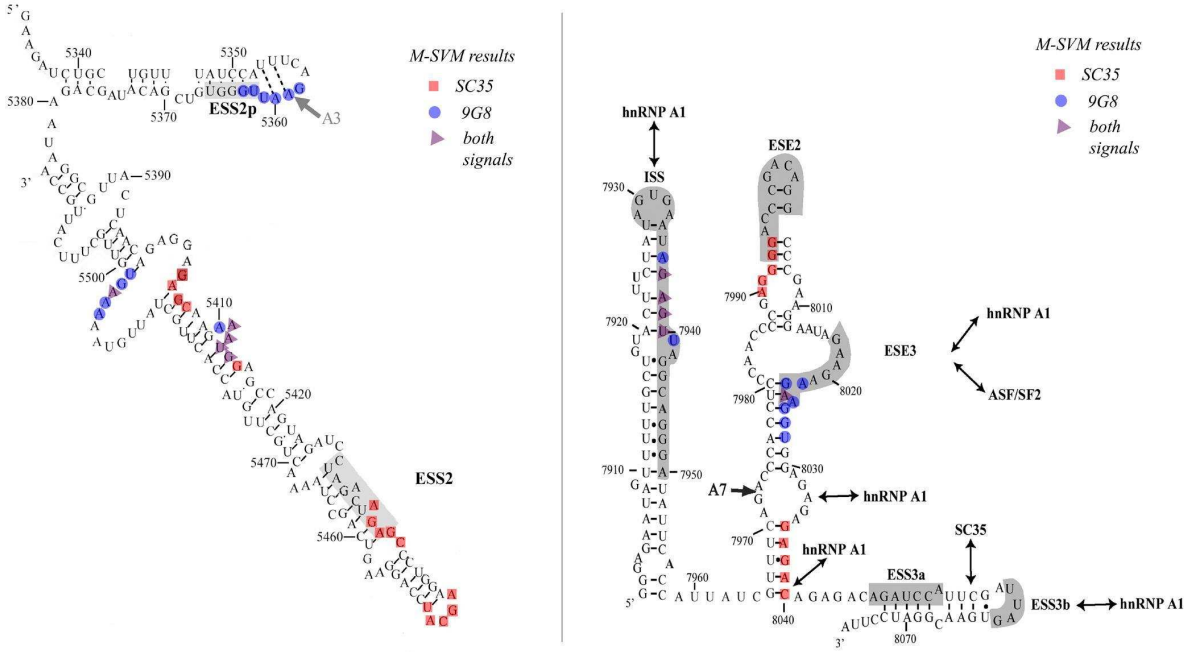

Figure 2: **M-SVM results for HIV-1 splicing acceptor A3 and A7 sites.**The structured regions of the HIV-1 RNA around A3 and A7 sites are represented ([12] [16]). The grey bars show the Exonic Splicing Silencer (ESS) and Exonic Splicing Enhancer (ESE). Red squares and blue circles mark window center positions of M-SVM signals for SC35 and 9G8, respectively. Purple triangles mark positions where both signals are found.

3.2.2 Results for the A7 site

This site is used to generate the tat and rev mRNA, producing the respective early phase regulatory proteins. For this site, we find two potential 9G8 binding sites in regions where hnRNPA1 has been shown to bind, the Intronic Splicing Silencer ISS and the Janus element also called ESE3 (for Exonic Splicing Enhancer 3, [16]). The ESE3 site being known to bind ASF/SF2 and hnRNPA1, there hence might be a three partners balance involved into the effect of this site reaching a level of complexity that begins to be difficult to deal with. It should be noticed that we were unable to retrieve a binding site for SC35 between ESS3a and b where it had been shown in [17]. Our short study of this site confirms the high level of complexity used to regulate its use. Considering that we did not study the binding properties of all the proteins known to be involved into alternative splicing, the authors would like to emphasize the interest there would be to try a formal modelling for these very complex phenomena to acquire a better undestanding of what is really going on during the virus life cycle.

A general comment about the M-SVM results is the global difficulty one has to get a sharp information about the boundaries of the potential interaction sequence given. Indeed, windows used in the training dataset are 18 nucleotides-long. When a 5 nucleotides-long hit is found this doesn't mean that those nucleotides are the target but that they are in the window of the target. This observation highlights the extreme care with which one has to take the results and confirms the need for experimental validation.

4 Discussion and perspectives

The quality of the results we obtained on the highly documented HIV-1 virus genome based on SELEX/M-SVM analysis are very encouraging when compared to independent experimental studies. The `grappe` tool is very useful to make an integrated analysis. To date, KOALAB software offers the opportunity to see and compare the results of the algorithmic and statistical learning methods in a same graphical output but it is also of interest to consider the possibility of real integration of the results from the two methods. An intermediary way of searching nucleic acid-protein interaction sequences is to use probabilistic models such as HMM that have been developed in the ESE finder tool. It would be at least of interest to confront all those techniques in a single study to find out whether an integration is of use or not.

Moreover, some sites were found by the M-SVM inside known splicing activating regions for which no further information about the involved proteins was available so far. This may lead to new hypothesis hence orienting the research of the biologist. The method could be further improved by adding SELEX data for other regulatory proteins. Our results clearly highlight the complementarity between the two methods employed and validates KOALAB method as a posteriori experimental results came and confirmed some of our hypothesis.

KOALAB has been introduced as a tool that helps experimental biologists to discover motifs. It now integrates an algorithmic and a statistical learning methods that can be confronted or used separately, depending on the heterogeneity of the motifs to be searched.

A classical intermediate method corresponds to the use of probabilistic models such as HMMs (Hidden Markov Models). In the case of alternative splicing regulation, HMM models have been developed for some SR proteins from SELEX data as well into a software called ESE finder (for Exonic Splicing Enhancer finder [3]). A possible extension of KOALAB would be to add such a method to be able to cover the full range of available methods for motif discovery. The user could make the choice of using only one or two best fitted methods according to the data and the biological purpose or, if possible, to use the whole repertoire of methods in order to confront their results throughout a single graphical interface.

After interpretation, one is able to identify parts of the genome that are potentially interesting in understanding alternative splicing regulation or any other problem, but still have to be validated experimentally. Furthermore, the resulting data can be used to formulate hypotheses on alternative splicing regulation which can be integrated into a formal model [7] that could lead to new hypotheses regarding global splicing regulation which could be experimentally tested.

Moreover, since the input of a SVM is not restricted to a real-valued vector, this machine could be extended to handle more complex data such as secondary structure information. Our software combines multi-class support vector machines with a discrete method to discover and validate motifs. In future work, KOALAB will integrate a variety of methods for biologists ranging from an algorithmic to a statistical learning approach including probabilistic automata. Following the idea that a computational method should be dedicated to a specific biological problem, KOALAB is designed for motif discovery fitting the biological purpose. Combining experimental analysis and computational tools is the key for providing an efficient motif discovery method.

Acknowledgement: The authors thank Y. Guermeur for M-SVM development, Y. Guermeur, C. Branlant and A. Bockmayr for their helpful comments during the writing of this paper and J. Stevenin for providing SELEX data.

References

[1] Bilodeau, P. S., Domsic, J. K., Mayeda, A., Krainer, A. R., Stoltzfus, C. M. (2001) RNA splicing at human immunodeficiency virus type 1 3' splice site A2 is regulated by binding of hnRNP A/B proteins to an exonic splicing silencer element. *J Virol.*, **75**(18), 8487–8497.

[2] Cai, C. Z., Wang, W. L., Sun, L. Z., Chen, Y. Z. (2003) Protein function classification via support vector machine approach. *Math Biosci*, **185**(2),111–122.

[3] Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., Krainer, A. R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**(13), 3568–3571.

[4] Cavaloc, Y., Bourgeois, C. F., Kister, L., Stevenin, J. (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA*, **5**(3), 468–483.

[5] Cortes, C., Vapnik, V. (1995) Support-Vector Networks *Machine Learning*,**20**(3), 273–297.

[6] Cristianini, N., Shawe-Taylor, J. (2000) An Introduction to Support Vector Machines and other kernel-based learning methods. *Cambridge University Press*

[7] Eveillard, D., Ropers, D., de Jong, H., Branlant, C., Bockmayr, A. (2003) Multiscale modeling of alternative splicing regulation. In *Computational Methods in Systems Biology (CMSB'03)*, Springer LNCS **2602**, 75–87.

[8] Graveley, B. R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6** (9), 1197–1211.

[9] Guermeur, Y., Elisseeff, A., Paugam-Moisy, H. (2000) A new multi-class SVM based on a uniform convergence result. *IJCNN'00*, **IV**, 183–188.

[10] Guyon, I., Boser, B. E., Vapnik, V. (1992) Automatic Capacity Tuning of Very Large VC-Dimension Classifiers. *NIPS*, 147–155. *Theoretical Computer Science*, **178**, 129–154.

[11] Hope, T. J. (1999) The ins and outs of HIV Rev. *Arch Biochem Biophys*, **365** (2), 186–191.

[12] Jacquenet, S., Ropers, D., Bilodeau, P., S., Damier, L., Mougin, A., Stoltzfus, C.,M., Branlant, C. (2001) Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing. *Nucleic Acids Res.*, **29**(2), 464–478.

[13] Kucherov, G. & Rusinowitch, M. (1997) Matching a set of strings with variable length don't cares. *Theoretical Computer Science*, **178**, 129–154.

[14] Liu, H. X., Zhang, R. S., Luan, F., Yao, X. J., Liu, M. C., Hu, Z. D., Fan, B. T. (2003) Diagnosing breast cancer based on support vector machines. *J Chem Inf Comput Sci.*, **43**(3), 900–907.

[15] Liu, H. X., Chew, S. L., Cartegni, L., Zhang, M.Q., Krainer, A. R. (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol.*, **20**(3), 1063–1071.

[16] Marchand, V., Mereau, A., Jacquenet, S., Thomas, D., Mougin, A., Gattoni, R., Stevenin, J., Branlant, C. (2002) A Janus splicing regulatory element modulates HIV-1 tat and rev mRNA production by coordination of hnRNP A1 cooperative binding. *J Mol Biol.*, **323**(4), 629–652.

[17] Mayeda, A., Badolato, J., Kobayashi, R., Zhang, M. Q., Gardiner, E. M., Krainer, A. R. (1999) Purification and characterization of human RNPS1: a general activator of prem-RNA splicing. *EMBO J*, **18** (16), 4560–4570.

[18] Qian, J., Lin, J., Luscombe ,N. M., Yu, H., Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*,**19**(15), 1917–1926.

[19] Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J., Mermod, N. & Bucher, P. (2002) Higth-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnology*, **20**, 831–835.

[20] Sun, Y. F., Fan, X. D., Li, Y. D. (2003) Identifying splicing sites in eukaryotic RNA: support vector machine approach. *Comput. Biol. Med.*, **33**(1), 17–29.

[21] Tuerk, C. & Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

[22] Vapnik, V. (1992) Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems 4*, 831–838.