SVM2Motif—Reconstructing Overlapping Sequence Motifs by Mimicking an SVM Predictor

1 Introduction

Major technological advances in sequencing techniques within the past decade have facilitated a deeper understanding of the mechanisms underlying the functionality and evolution of molecular processes. Considering the sheer size of many genomes, it comes, however, at the expense of an enormous amount of data that demands for automatic and computationally efficient methods in genomic discrimination tasks (see, for instance, ENCODE project).

One of the most accurate approaches to predictive sequence analysis is the support vector machine (SVM) [2] along with the use of a weighted-degree (WD) kernel (e.g., [1]). The WD kernel compares two nucleotide or amino-acid sequences x and x' of length L by dividing them into all subsequences of length $k \in \mathcal{K} \subset \mathbb{N}$, where $\bar{k} := \max_{k \in \mathcal{K}} k \leq L$, and then counting the number of matching subsequences—the so-called *positional oligomers* (POs). A WD-kernel SVM then employs the simple, linear model $s(x) := \langle w, \Phi(x) \rangle$, which, denoting $x[i]^l := (x_i, \ldots, x_{i+l-1})$, can be written as

$$s(x) = \sum_{l=1}^{k} \sum_{i=1}^{L-l+1} w_{(x[i]^l, i)} \,. \tag{1}$$

Unfortunately, due to its black-box character, biological factors underlying the SVM's prediction such as transcription factor binding sites or start sites—often called *motifs*—are largely unknown and cannot be easily extracted from the SVM classifier. An important first step towards the identification of motifs was proposed in [4, 7]: the concept of *positional oligomer importance matrices* (POIMs). POIMs assign each PO y of length k starting at position j with an importance score $\text{POIM}_{j,y} \sim \mathbb{E}[s(\mathcal{X}) | \mathcal{X}[j]^k = y]$. It allows visualization of the PO's significance. An important property of POIMs is that they take the overlaps of the POs at different positions and of different lengths into account.

Although being a major step towards the explanation of trained SVM models, POIMs suffer from two problems: a) the conceptual problem that one has to analyze multiple and combinations of POs to understand the motif associated with the biological process and b) the computational problem that the size of POIMs grows exponentially in the length of the motif (see Figure 1). This renders their computation feasible only for rather small motif sizes ($k \leq 12$) and hampers manual inspection necessary to determine candidate motifs already for rather small motif sizes ($k \approx 5$).

In this work, we tackle the problem of obtaining motifs from the output of an SVM via the use of POIMs from a different perspective. In a nutshell, our approach is the other way round (!): we propose a probabilistic framework to reconstruct, from a given motif, the POIM that is the most likely to be generated by the motif. By subsequently minimizing the reconstruction error with respect to the given SVM-generated POIM, we can in fact optimize over the motif in order to find the one that is the most likely to have generated the POIM at hand. The latter poses a substantial numerical challenge due to the extremely high dimensionality of the feature space. Figure 2 illustrates our approach.

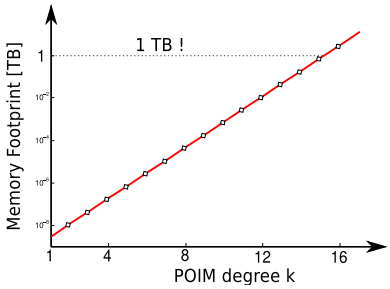


Figure 1: Memory requirements of POIMs.

The main contributions of this paper can be summarized as follows:

1. Putting forward the work of [7] on positional oligomer importance matrices (POIMs), we present a novel framework for extracting relevant motifs underlying the prediction of a WD-kernel SVM.
2. To deal with the sheer unfeasibly large size of the feature space associated with the WD kernel, we propose an efficient numerical framework containing numerous speed-ups such as logical bit-shift operations and progressive sequence decomposition techniques, as well as we provide a free open-source implementation thereof, which will be made available.
3. Our approach is able to even find *overlapping* motifs consisting of up to hundreds of nucleotides, while previous approaches are limited to comparably short and contiguous motifs.
4. We demonstrate the efficiency and efficacy of our approach on synthetic data sets as well as a *human* splice data set. To access the performance of our approach, we compare the so-obtained results to known motifs from the JASPAR database [5]. Additional evidence from further experiments on transcription start site (TSS) data, which are currently in progress, are expected to be available in time for presentation at the workshop.

2 Methodology

Let $\Sigma = \{A, C, G, T\}$ be the DNA alphabet and let $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_L) \sim \mathcal{U}(\Sigma^L)$ be a random vector that is uniformly distributed over Σ^L for some $L \in \mathbb{N}$. Let $y \in \Sigma^k$ be a k-mer and let j be a valid position of y in $\mathcal{X}$, i.e., $j \in \{1, \ldots, L-k+1\}$.

2.1 POIMs

To measure, for a given SVM hypothesis w, the contribution of the positional k-mer (y, j) to the SVM output function $s(\cdot)$ as defined in (1), Sonnenburg et al. [7] propose to consider the quantity

$$Q^k_{y,j} := \mathbb{E}[s(\mathcal{X})|\mathcal{X}[j]^k = y] - \mathbb{E}[s(\mathcal{X})]\,. \tag{2}$$

The above equation (2) can be interpreted as follows: suppose the positional k-mer (y, j) indeed had a substantial impact on the function $s(\cdot)$, then its expected value would be likely to increase when conditioning on the occurrence of y at position j in $\mathcal{X}$. Motivated by this reasoning, [7] study the structure of the set

$$Q := \{Q^l_{y,j} : \; l = 1, \ldots, k, \; y \in \Sigma^l, \; j = 1, \ldots, L-l+1\}\,,$$

which they refer to as *positional oligomer importance matrices* (POIMs). An example of a POIM is shown in Figure 3, where the highest scoring positional oligomer is shown in red. POIM sets can be analysed, e.g., by means of their *differential POIM*, as proposed in [8]. The differential POIM D is defined as a matrix with entries

$$D_{l,j} := \bar{q}_{l,j} - \max(\bar{q}_{l-1,j}\,,\, \bar{q}_{l-1,j+1})$$

where $\bar{q}_{l,j} := \max_{y \in \Sigma^k} \left|Q^l_{y,j}\right|$ and for $l = 2, \ldots, k$, $j \in \{1, \ldots, L\}$ and 0 otherwise. The entry $D_{l,j}$ can be interpreted as an overall score for the general importance of the oligomers of length l at position j so that the differential POIM can be put to good use for estimating the position and length of a motif. However, both, the vanilla POIM and the differential POIM, are, per se, unable to specify a motif's sequence.

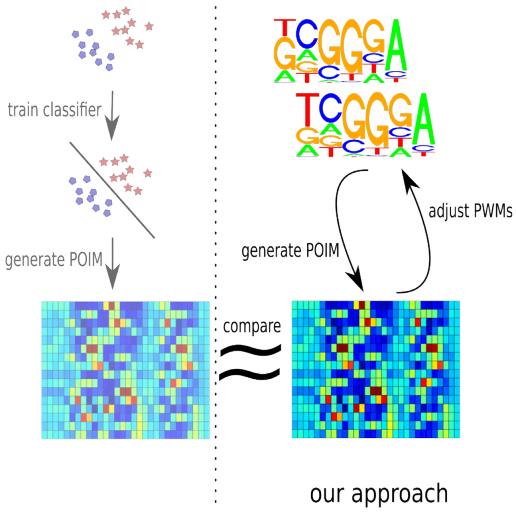


Figure 2: Illustration of the proposed approach: extracting a motif (top right) from a trained SVM model (top left) by approximating the SVM POIM (bottom left) by a POIM generated from a set of motifs (bottom right).

2.2 motifPOIMs

We define a *probabilistic positional motif* (PPM) as a tuple $m_k := (r, \mu, \sigma)$, consisting of a *positional weight matrix* (PWM) $r \in \mathbb{R}^{4\times k}$ anchored at a starting position $\mu \in \mathbb{R}$ that is subject to some uncertainty encoded by the standard deviation $\sigma \in \mathbb{R}$. In order to approximate the SVM weight vector w occurring in (1), we define a vector v with entries

$$v_{(z,i)}(m_k) := \frac{1}{\sqrt{2\pi}\sigma} exp\Big(- \frac{(i-\mu)^2}{2\sigma^2}\Big) \prod_{l=1}^{k} r_{z_l,l}\,,$$

for any positional k-mer $(z, i) \in \Sigma^k \times \{1, \ldots, L-k+1\}$. Consequently, in analogy to (1), we define a function

$$\tilde{s}(x; m_k) := \sum_{i=1}^{L-k+1} v_{(x[i]^k,i)}(m_k)\,, \tag{3}$$

so that, given a set of motifs m_k, $k \in \mathcal{K}$, where $\mathcal{K} \subset \mathbb{N}$ is the set containing the motif lengths, we can construct a motif-based approximation of the POIM consisting of the entries

$$R_{y,j}(m_k) := \mathbb{E}[\tilde{s}(\mathcal{X}|m_k)|\mathcal{X}[j]^k = y] - \mathbb{E}[\tilde{s}(\mathcal{X}|m_k)], \quad k \in \mathcal{K}\,,$$

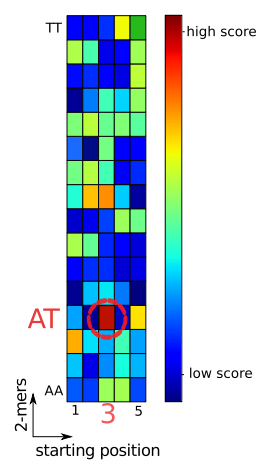


Figure 3: Illustration of a POIM.

which we call *motifPOIM* and denote by R. Our overall aim is, by optimizing over the motifs m_k, to approximate the original POIM by the motifPOIM $R(m_k)$ (cf. Figure 2 from the introduction for an illustration of the proposed workflow). Note that we can in fact compute motifs for almost arbitrarily long motifs/PPMs m_k by modelling a PPM of length

$\tilde{k} \leq k$ as a number of $\mathcal{D} := k - \tilde{k} + 1$ many overlapping SubPPMs $\tilde{m}_d(m_k, \tilde{k}) := (\tilde{r}, \tilde{\mu}, \sigma)$, $\forall\, d = 0, \ldots, D-1$, with $\tilde{\mu} := \mu + d$ and the $\tilde{k}$ column sub-matrix $\tilde{r}$ of r starting at column d.

The core idea of our numerical approach now is to find a set of motifs m_k, $k \in \mathcal{K}$, such that $\sum_{d=0}^{D-1} R_{y,j}(\tilde{m}_d(m_k, \tilde{k})) \approx Q^{\tilde{k}}_{y,j}$ for most oligomers (y, j) as given by the quadratic distance $\Delta^k_{y,j}$ between each entry $\Delta^k_{y,j}(m_k) = \big(\lambda_k \sum_{d=0}^{D-1} R_{y,j}(\tilde{m}_d(m_k, \tilde{k})) - Q^{\tilde{k}}_{y,j}\big)^2$, where $\lambda_k \geq 0$ is the weight associated with the motif m_k. This is encoded in the following primal optimization problem (the corresponding dual representation is omitted for space constraints):

Problem 1 (PRIMAL MOTIFPOIM OPTIMIZATION PROBLEM)**.**

$$\begin{aligned} \inf_{m_k, \lambda_k} \quad & \frac{1}{2} \sum_{k \in \mathcal{K}} \sum_{y \in \Sigma^k} \sum_{j=1}^{L} \Delta^k_{y,j}(m_k) \\ \text{s.t.} \quad & \forall k \in \mathcal{K}: \;\; 0 < \sigma_k \leq k, \;\; 1 \leq \mu_k \leq L - k + 1, \;\; 0 \leq \lambda_k \\ & \forall k, s, t: \;\; 0 \leq r^k_{s,t}\,, \end{aligned} \tag{P}$$

We solve the above problem, we furthermore exploit that the computation of the objective function can be considerably speed-up by the use of bit-shift operations as indicated by the following theorem:

Theorem 2. *Let $\mathcal{A}$ the dependency matrix of y. Let $\mathcal{C}^j(y|m_k)$ be a matrix over $\mathcal{I}_y(k)$ of size $\Sigma^k \times (2(k-1)+1)$ containing the weights of all positional oligomers in $\mathcal{A}(y)$, i.e., $\mathcal{C}^j_{z,i}(y|m_k) = v_{(z,i+j-k+1)}(m_k)$ if $1 \leq j + i - k \leq L$ and 0 otherwise. Then it holds:*

$$R_{y,j}(m_k) = \langle \mathcal{A}(y), \mathcal{C}^j(y|m_k) \rangle\,.$$

Proof. Omitted for space constraints. □

3 Empirical Evaluation

In the computational experiments, we employ the SHOGUN machine-learning toolbox [6] for SVM training and POIM computation. We also deploy the differential POIM approach to estimate the number of motifs as well as their length and starting position, all which serve as an initialization of Problem 1, which we solve by the L-BFGS-B Algorithm as implemented by [3]. The parameters are set to default values: regularization constant $C = 1$, WD-Kernel degree $d = 20$, and a maximal POIM degree $k = 7$.

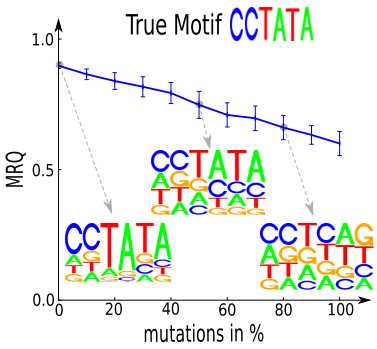
3.1 Controlled Experiments

We generate three synthetic data sets (will be made available), addressing the following scenarios:

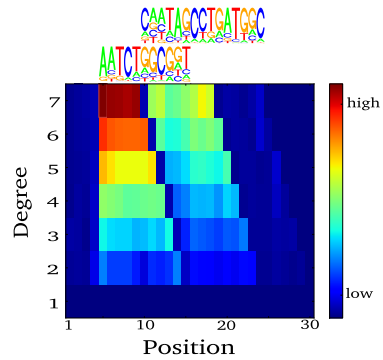
1. MUTATION. We generate a sample consisting of 10,000 random DNA sequences of length 30, where we insert the motif CCTATA into the positive examples. The motif is camouflaged by, for each example, flipping any of the six motif nucleotides, each with probability $p \in [0, 1]$. The data set is realized for various values of p. See Figure 4a.
2. OVERLAPPING. As above, but we insert two overlapping unmutated positional oligomers (AATCTGGCGGT,5) and (CAATAGCCTGATGGC,10); each PO is (exclusively) inserted into 50% of the training examples. See Figure 4b.
3. LONG. As above, inserting a single unmutated motif (TCGGATCGGAT...) of length 200 into sequences of length 400. See Figure 4c.

We measure the motif reconstruction quality (MRQ) by the JASPAR score [5], defined as

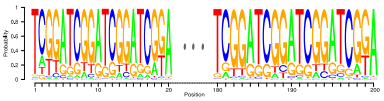
$$\text{MRQ} = \sum_{l=1}^{k} \left(\frac{1}{k} - \frac{1}{2k} \sum_{c \in \{A,C,G,T\}} (t_{cl} - r_{cl})^2 \right).$$



(a) MUTATION



(b) OVERLAPPING



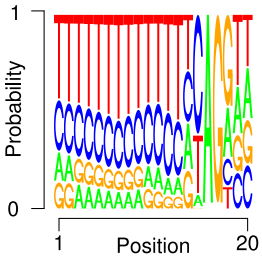
(c) LONG

Figure 4: Results of the experiments on synthetic data sets.

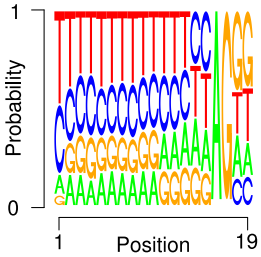
Results The results for the three considered scenarios are shown in Figure 4 from top to bottom. For the MUTATION data set, we can observe that, up to a mutation level of 50%, we correctly identify the true underlying motif as being the sequence with the highest probability in the PWM. In the worst case of a very high mutation, the quality of reconstruction decreases but our approach can still roughly indicate conforming parts of the true motif (the first half of the motif identified correctly). For the OVERLAPPING data set, the differential POIM indicates two accumulations with high scores, from which we observe starting positions of the two different, overlapping motifs. Although the two motifs are overlapping, both motifs are identified correctly. For the LONG data set, due to the immense number of nucleotides in the motif, we divide the corresponding motifPOIM into 10 smaller conforming parts, in each searching for a motif of smaller length ($k = 20$). We observe that our algorithm accurately assembles the considerably long motif.

3.2 Real-World Experiments on Human Splice Data

We downloaded a *human* splice data set from `http://www.fml.tuebingen.mpg.de/raetsch/projects/lsmkl`. For verifying our results we use the JASPAR database,[1] which contains a collection of important DNA motifs for splice sites as well as transcription start sites. Non-polymorphic loci are dealt by manual post-processing. The results are shown in Figure 5. The differential POIM indicates a motif of length 19 starting at position 85. Comparing the motif found by our approach with the true JASPAR motif, we observe a striking accordance as evidenced by a considerably high JASPAR consensus score of 98.39. Note that, for example, a completely random sequence (19 uniformly drawn nucleotides) has an average consensus of 89.31, which is greatly exceeded by our result.



(a) True JASPAR motif



(b) Our approach (JASPAR score of 98.39)

Figure 5: Experimental results for the human splice site detection experiment.

4 Conclusion

Identifying discriminative motifs underlying the functionality and evolution of molecular processes is a major challenge in computational biology. Putting forward the work of [7] on positional oligomer importance matrices (POIMs), we have developed a new methodology, entitled SVM2Motif, to automatically extract the truly relevant motifs from an SVM predictor—even if the motifs are overlapping or obfuscated by mutation. We have considered the candidate motifs as free parameters in a hierarchical probabilistic model, a task which can be phrased as a non-convex optimization problem. The exponential dependence of the POIM on the oligomer length poses a major numerical challenge, which we address by an efficient optimization framework implementing logical bit-shift operations and progressive sequence decomposition techniques. We demonstrate the efficacy of our approach on several synthetic data sets as well as a real-world *human* splice site data set. Additional evidence from further experiments on transcription start site (TSS) data, which are currently in progress, are expected to be available in time for presentation at the workshop. The source code will be made available (URL omitted to maintain anonymity).

References

[1] A. Ben-Hur, C.-S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. Support vector machines and kernels for computational biology. *PLoS Comput Biology*, 4(10):e1000173, October 2008.

[2] B.E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In David Haussler, editor, *COLT*, pages 144–152. ACM, 1992.

[3] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(3):503–528, 1989.

[4] G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. Müller, R.-J. Sommer, and B. Schölkopf. Improving the caenorhabditis elegans genome annotation using machine learning. *PLoS Comput Biol*, 3(2):e20, Feb 2007.

[5] A. Sandelin, W. Alkema, P.G. Engström, W. W. Wasserman, and B. Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(Database-Issue):91–94, 2004.

[6] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. De Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, 11:1799–1802, 2010.

[7] S. Sonnenburg, A. Zien, P. Philips, and G. Rätsch. POIMs: positional oligomer importance matrices — understanding support vector machine based signal detectors. *Bioinformatics*, July 2008. (received the Outstanding Student Paper Award at ISMB 2008).

[8] A. Zien, P. Philips, and S. Sonnenburg. Computing Positional Oligomer Importance Matrices (POIMs). Research Report; Electronic Publication 2, Fraunhofer Institute FIRST, December 2007.

[1] `http://jaspar.genereg.net`