KIRMES: Kernel-based Identification of Regulatory Modules in Euchromatic Sequences

Sebastian J. Schultheiss,[1,2] Wolfgang Busch,[2] Jan U. Lohmann,[2]
Oliver Kohlbacher,[3] and Gunnar Rätsch[1]

[1]Friedrich Miescher Laboratory of the Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany

[2]Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany

[3]University of Tübingen, Wilhelm Schickard Institute for Computer Science,
Division for Simulation of Biological Systems, Sand 14, 72076 Tübingen, Germany

Abstract:
Motivation: Understanding transcriptional regulation is one of the main challenges in computational biology. An important problem is the identification of transcription factor binding sites in promoter regions of potential transcription factor target genes. It is typically approached by position weight matrix-based motif identification algorithms using Gibbs sampling or heuristics for extending seed oligos. Such algorithms succeed in identifying single, relatively well conserved binding sites, but tend to fail when it comes to the identification of combinations of several degenerate binding sites as those often found in *cis*-regulatory modules.
Results: We propose a new algorithm that combines the benefits of existing motif finding with the ones of Support Vector Machines (SVMs) to find degenerate motifs in order to improve the modeling of regulatory modules. In experiments on microarray data from *Arabidopsis thaliana* we were able to show that the newly developed strategy significantly improves the recognition of transcription factor targets.
Availability: The PYTHON source code (open source–licensed under GPL), the data for the experiments and a web-service are available at `http://www.fml.mpg.de/raetsch/projects/kirmes`.
Contact: sebi@tuebingen.mpg.de

1 Introduction

One of the most important problems in understanding transcriptional regulation is the prediction of transcription factor target genes based on their promoter sequence. A transcription factor binding site (TFBS) is a short sequence segment ($\approx$10 bp) located near a gene's transcription start site (TSS) and is recognized by respective transcription factors (TFs) for gene regulation [GL05]. TFBSs recognized by the same TF usually show a conserved pattern, which is often called a TF binding motif (TFBM) [GL05]. Such TFBMs are typically identified by considering overrepresented motifs in promoter sequences of a set of genes that is enriched with targets for a specific transcription factor. The simplest approaches include the identification of overrepresented oligomers relative to a background model [BE94]. More sophisticated models include Gibbs-sampling methods [LAB$^+$93] that try to identify position weight matrices[SSGE86] characterizing binding sites in the candidate promoter sequences [Sto00].

Although these methods have been very successful for bacterial and yeast genomes, their success was limited in higher eukaryotes for which TFBMs are often degenerate and the search space is considerably larger. While some recent techniques have improved the

state-of-the-art, they all tend to fail if the motif is defined only weakly or in the context of other motifs. "Despite these challenges, there are two possible redeeming factors: (i) many eukaryotic genomes have been or are being sequenced, and comparative genomic analysis can be extremely powerful; and (ii) most eukaryotic genes are controlled by a combination of factors with the corresponding binding sites forming homotypic or heterotypic clusters known as '*cis*-regulatory modules' (CRMs)" [GL05].

In this work we developed novel methods that are able to classify genes as being either TF targets or not, based on the presence of motifs and features capable of describing CRMs. This is done by a two-step procedure. We first used *de novo* motif finding tools or known motif databases like TRANSFAC [MFG$^+$03] or JASPAR [SAE$^+$04] to identify a set of potential motifs. Then we used SVMs employing a newly developed kernel that is capable of capturing information about the motifs and their relative location to classify promoter sequences. Additionally, we demonstrate the potential of our approach to exploit conservation information to improve the classification performance.

Most previous approaches for discovering CRMs are based on the identification of motifs and their co-occurrences (*e.g.* [FSKB08, ST02]). Other approaches exploit site-clustering information with *de novo* motif discovery to build rules discriminating modules that preserve the ordering of motifs (*e.g.* [SS05]). Finally, [YTI$^+$98] suggested to use Hidden Markov Models to represent CRMs and [GL05] developed a Monte Carlo method and dynamic programming approach to screen motif candidates. The main difference between our approach and most previous approaches is that we use discriminative methods that allow us to model the TFBS' more accurately. In particular, instead of using zeroth-order inhomogeneous Markov chains, we use Support Vector kernels to model higher order sequence information around putative TF binding sites.

The paper is organised as follows: We start Section 2.1 by describing the basic methodology of classifying sequences with Support Vector Machines using standard sequence kernels. It is followed by a detailed explanation of the main idea of this work in Section 2.2 for combining *de novo* motif finders with state-of-the-art motif modeling. In Section 3 we outline a problem derived from *A. thaliana* microarray expression experiments where certain transcription factors are over- or under-expressed. In our experiment we first illustrate that the straightforward approaches cannot achieve reasonable results, while the newly developed methods are able to drastically improve the target gene recognition performance.

2 Methods

2.1 Sequence Classification with Support Vector Machines

Support Vector Machine (SVMs) are a well-established machine learning method introduced by Boser, Guyon, and Vapnik [BGV92] to solve classification tasks frequently appearing in computational biology and many other disciplines. Typical examples are the classification of tumor images or gene expression measurements, the detection of biological signals in DNA, RNA or protein sequences as well as the recognition of hand-written digits or faces in images. SVMs are widely used in computational biology due to their high accuracy, their ability to deal with high-dimensional data, and their flexibility in modeling diverse sources of data [MMR$^+$01, SS02, STV04, Nob06].

The domain knowledge inherent in the classification task is captured by defining a suitable *kernel function* $k(\boldsymbol{x}, \boldsymbol{x}')$ computing the similarity between two examples $\boldsymbol{x}$ and $\boldsymbol{x}'$. This strategy has two advantages: the ability to generate non-linear decision boundaries using methods initially designed for linear classifiers; and the possibility to apply a classifier to data that have no obvious vector space representation, for example, DNA/RNA or protein sequences as well as structures [BOS+08].

Spectrum Kernel Given two example sequences $\boldsymbol{x}$ and $\boldsymbol{x}'$ over the alphabet Σ, a simple way to compute the similarity is to count the number of co-occurring oligomers of fixed length ℓ. This idea is realized in the so-called *spectrum kernel* that was first proposed for classifying protein sequences [LEN02]: $k_\ell^{\text{spec}}(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi_\ell^{\text{spec}}(\boldsymbol{x}), \Phi_\ell^{\text{spec}}(\boldsymbol{x}') \rangle$, where $|\Sigma|$ is the number of letters in the alphabet. Φ_ℓ^{spec} is a mapping of the sequence $\boldsymbol{x}$ into a $|\Sigma|^\ell$-dimensional feature-space. Each dimension corresponds to one of the $|\Sigma|^\ell$ possible strings s of length ℓ and is the count of the number of occurrences of s in $\boldsymbol{x}$. This kernel is well-suited to characterize sequence similarity based on oligos that appear in both sequences– independent of their position.

If the classification of promoter sequences of genes as transcription factor targets would be solely based on binding to specific oligos, then the spectrum kernel appears to be a reasonable choice. If the motif is less conserved, then allowing for mismatches or gaps can be beneficial [LEWN03]. Note that this kernel is (by design) incapable of recognizing positional preferences TFs, and thus TFBSs, might have relative to the transcription start or among each other.

Weighted Degree Kernel Another kernel, the so-called *Weighted Degree Kernel* (WD) was proposed in [RS04, SRR07]. It computes the similarity of sequences of fixed length L by considering the substrings up to length ℓ starting at each position l separately:

$$k_\ell^{\text{wd}}(\boldsymbol{x}, \boldsymbol{x}') \quad = \quad \sum_{l=1}^{L} \sum_{d=1}^{\ell} \frac{\beta_d}{L} \mathbf{I}\left(\boldsymbol{x}_{[l:l+d]} = \boldsymbol{x}'_{[l:l+d]}\right) \text{ where } \beta_d = 2\frac{\ell - d + 1}{\ell^2 + \ell}, \qquad (1)$$

and $\boldsymbol{x}_{[l:l+d]}$ is the substring of length d of $\boldsymbol{x}$ at position l [RS04, SRR07].

In the WD kernel, only oligos appearing at the same position in the sequence contribute to the similarity of two sequences. The *WD kernel with shifts* [RSS05] is an extension of the WD kernel allowing some positional flexibility of matching oligos:[1]

$$k_{\ell,S}^{\text{wds}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{l=1}^{L} \sum_{d=1}^{\ell} \sum_{\substack{s=0 \\ s+i \le l}}^{S} \frac{1}{2dL(S+1)} \left(\mathbf{I}\left(\boldsymbol{x}_{[l+s:l+d+s]} = \boldsymbol{x}'_{[l:l+d]}\right) + \mathbf{I}\left(\boldsymbol{x}_{[l:l+d]} = \boldsymbol{x}'_{[l+s:l+d+s]}\right) \right) \qquad (2)$$

It considers oligomers up to length d, and allows them to be shifted up to S positions, starting from i, in the input sequences. This kernel is better suited for motifs with indels or at varying positions (see *e.g.* [RSS05, SSP+07]

[1] The *locality improved* and *oligo* kernel [ZRM+00, MTMM04] achieve a similar goal in a slightly different way.

2.2 Extensions

In this section we extend the WD kernel in two different ways: First, we consider an extension to use conservation information. Second, given a list of potential motifs we propose a new kernel that integrates information on the motif sequences with the information about their co-occurrence with the aim to characterize regulatory modules.

WD Kernel with Conservation Information To include conservation information, we extended the WDS kernel with a term to multiply the score of the local matches of an oligo of length d at position i with a quantity that depends on its conservation. We propose to use the average conservation of the oligo in pre-generated alignments of sequences from G other organisms:

$$\gamma^A_{d,i,\boldsymbol{x}} = 1 + \frac{A}{d} \sum_{g=1}^{G} \sum_{j=0}^{d} \mathbf{I}(\boldsymbol{x}_{i+j} = \boldsymbol{x}^g_{i+j}), \tag{3}$$

where $\boldsymbol{x}^g$ is the sequence of the syntenic regions in the genome of organism $g = 1, \ldots, G$ and $A < 0$ is a parameter allowing one to control the importance of the conservation. The fact that we add 1 means we only value an existing alignment positively, but do not further punish the absence of an alignment. All results shown were obtained with the setting of $A = 1$. Using this definition of a conservation score we can now define the *weighted degree with shifts and conservation* (WDSC):

$$k^{\text{wdsc}}_{\ell,S,A}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{l=1}^{L} \sum_{d=1}^{\ell} \sum_{\substack{s=0 \\ s+i \le l}}^{S} \frac{\gamma_{d,i,\boldsymbol{x}} \gamma_{d,i,\boldsymbol{x}'}}{2d(s+1)} \left(\mathbf{I}\left(\boldsymbol{x}_{[l+s:l+d+s]} = \boldsymbol{x}'_{[l:l+d]}\right) + \mathbf{I}\left(\boldsymbol{x}_{[l:l+d]} = \boldsymbol{x}'_{[l+s:l+d+s]}\right) \right)$$

A Kernel for Regulatory Modules

Suppose we are given a set of M motifs $\mathcal{M}_m$, $m = 1, \ldots, M$ that may either come from a database or from a *de novo* motif detection method. Such motifs are often represented in a way that one can easily scan a given sequence for occurrences of the motif (*e.g.* as PWMs). In a preprocessing step we compute the best-matching position $p_{m,\boldsymbol{x}}$ of each motif $\mathcal{M}_m$ in all considered sequences $\boldsymbol{x}$. In case of PWMs, the PWM score and in case of oligo-based motifs the Hamming distance may be used to decide which position in the sequence matches best.[2]

The main idea of the kernel that we propose is to represent an input sequence $\boldsymbol{x}$ by the set of sequences $\boldsymbol{x}_m := \boldsymbol{x}_{[p_{m,\boldsymbol{x}}-w, p_{m,\boldsymbol{x}}+w]}$ originating from the region of length $2w$ around the best motif match $p_{m,\boldsymbol{x}}$ of motif $\mathcal{M}_m$. Each sequence region $\boldsymbol{x}_m$ contributes independently to the similarity between two input sequences: $k_1(\boldsymbol{x}, \boldsymbol{x}') = \sum_{m=1}^{M} k(\boldsymbol{x}_m, \boldsymbol{x}'_m)$. This term characterizes the co-occurrence of a collection of motifs in two sequences $\boldsymbol{x}$ and $\boldsymbol{x}'$. The similarity is highest if all motifs appear in both sequences (in arbitrary order). We propose to use a position specific kernel, for instance the WDS kernel, to compute the similarity of the regions.

[2] For the kernel functions, all input vectors need to be of the same length. Therefore, in our method we have to choose the same number of matches per sequence for all motifs (1 in our case), regardless of the quality of the matches. Biologically, a threshold quality seems more intuitive, then several good matches would be considered and no match for sequences that don't contain the motif. However, a soft margin during training allows the algorithm to ignore some mislabeled data points without effects on generalization.

For the first part of the kernel, the position of the motif does not influence the similarity at all. In the second part of the kernel we try to capture the relative position of the best motif matches to each other and to the transcription start site. This is done by computing all pairwise distances between match positions of motifs: $v(\boldsymbol{x}) = (p_{1,\boldsymbol{x}} - p_{tss}, \ldots, p_{M,\boldsymbol{x}} - p_{tss}, p_{1,\boldsymbol{x}} - p_{2,\boldsymbol{x}}, \ldots, p_{i,\boldsymbol{x}} - p_{j,\boldsymbol{x}}, \ldots, p_{M-1,\boldsymbol{x}} - p_{M,\boldsymbol{x}})^\top$, for all $i \neq j = 1, \ldots, M$, where p_{tss} is the position of the transcription start site in the sequence. A simple way of computing the similarity between two such vectors is to use the RBF kernel (*e.g.* [SS02]): $k^{\mathrm{rbf}}(\boldsymbol{v}, \boldsymbol{v}') = \exp\left(\frac{\|\boldsymbol{v}-\boldsymbol{v}'\|^2}{\sigma}\right)$, where σ is a kernel hyper parameter to be found by model selection.

Having both parts of the kernel defined, the question remains of how to combine them. We propose to simply add both contributions: $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}') + k^{\mathrm{rbf}}(v(\boldsymbol{x}), v(\boldsymbol{x}'))$. Please note that if we add the two kernels, it amounts to concatenating the two feature spaces. If one would multiply the contributions of distances and motif-sequence similarity, then the kernel would be in some sense similar to the previously proposed oligo kernel [MTMM04].

2.3 KIRMES Pipeline

Below we describe an integrated PYTHON [Pyt07] pipeline, called KIRMES, using the previously described kernels to classify promoter regions of genes as transcription factor targets or not.[3] It assumes that the sequences of promoter regions are given in two sets: A set enriched with transcription factor targets (labeled positive) and a second set containing no or very few targets (labeled negative). Figure 1 shows an outline of the pipeline for the classification of promoter sequences based on microarray experiments (*cf.* Section 3.1).

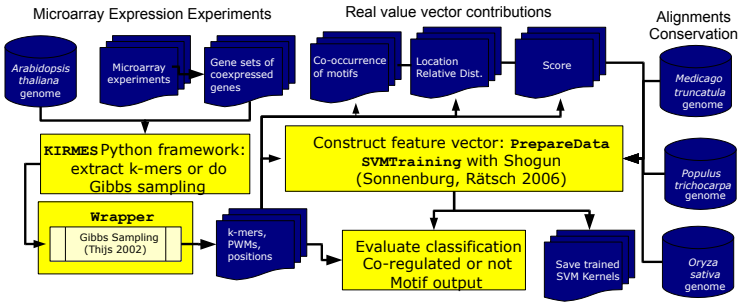

Figure 1: Workflow of KIRMES: The pre-processing step requires the genomic sequence and a set of genes that were measured to be coexpressed in microarray experiments. KIRMES extracts the k-mers and puts them into a vector along with the positional and conservation information and the score, as described in Section 2.2. The SVM is trained on a labeled data set of positives and negatives and can then be applied repeatedly on unlabeled testing datasets.

[3] Sequences considered in the set can be any part of the euchromatin of arbitrary length, *e.g.* upstream and downstream regions of a gene, intronic and exonic parts as well as untranslated regions (UTR) in the 3' or 5' direction. Good results can be obtained with a combination of upstream, UTR and intronic sequences. We used 2000 kb upstream of the translation start; in general longer sequences introduce more noise. Therefore, in organisms with shorter promoters a reduction would be beneficial for the signal to noise ratio.

Initial Motif Finding In a first step we used one of two methods to identify potential motifs. Initially, we used a common Gibbs sampling algorithm [LAB+93] called MOTIF-SAMPLER from the INCLUSIVE package [TMDS+02] that finds overrepresented motifs. To make sure we do not include motifs that are too common, we use several strategies: first, a background model for this organism; second, minimum occurrences were set to 15% or three genes of the set, whichever is more; third, one thousand random gene sets were generated and searched for motifs of the same length and determinacy. This was measured through the information content of the position frequency matrix of the motif, an output of the Gibbs sampling program.

Since this last step takes a significant amount of time depending on the length of the sequences, we searched for alternatives. We settled on one approach, where we count the occurrence of any oligomer of length six in positive sequences (*oligo-counting*). We select a subset of those oligomers that appear in at least 15% of all positive sequences. This simple strategy certainly leaves room for improvements, but our experiments in Section 3 illustrate that it already works rather well.

SVM Training We use the large scale machine learning toolbox SHOGUN [SRSS06] through its PYTHON interface. It provides implementations of all kernels described in this work and allows fast training using several different SVM implementations, *e.g.* SVM^{light} [Joa99].

Galaxy Web Service KIRMES is available publicly on our Galaxy webserver at `http://www.fml.mpg.de/raetsch/projects/kirmes`. Galaxy is an open-source, scalable framework for tool and data integration [GRH+05]: Users can upload their sequence files and KIRMES will classify the input gene set and return the names of the co-regulated genes in a list. This can be done for any regulatory region like promotors, introns, or even the whole chromatin of arbitrary length, and for any organism. To successfully use the positional information in promotor regions, it is a good idea to select the sequences in such a way that the translation start site is at the same position in each of them.

For this web service, the 6-mer enumeration strategy and the weighted degree kernel with shifts is used. The use of conservation information is not supported as it depends on the organism from which the sequences were obtained, it may not always be available and would require a significantly larger infrastructure. There is no upper limit on the amount of input sequences, but at least 5 sequences should be uploaded.

3 Experiments

We first describe a dataset which has been used to test the presented methods. The goal is to predict the expression change status of potential target genes for over-expressed transcription factors based on their promoter sequence.

3.1 Microarray Expression Data

We derive sets of co-expressed genes from microarray experiments performed with the commercial *Affymetrix GeneChip Arabidopsis ATH1* array. This chip is designed to measure transcript abundance of more than 20 000 genes of the model organism *Arabidopsis thaliana* [RHTT04].

The sets are obtained through a stringent analysis of expression change using the software GeneSpring [Agi06]. We labeled genes as co-expressed when they showed a four-fold change of expression in the experiment as compared to the control, and considered those genes not co-expressed if their levels remain the same, compared to the control, within a margin of 0.2 fold change. Thus we obtained sets of co-expressed genes.[4]

We used microarray data from two different experimental setups (*cf.* Appendix A.1 in the Supplementary Materials). The first setup uses leaves from wild type *Arabidopsis thaliana* plants exposed to medium at 38 °C *versus* leaves exposed to the same medium at room temperature, expression measurement taken one hour after exposure [BWS05]. The second setup uses inducible over-expression of *Arabidopsis* meristem regulators with the AlcR/AlcA system. Plants harboring 35S::AlcR/AlcA::GOI (GUS control, LEAFY, SHOOTMERISTEMLESS, WUSCHEL) constructs were grown in continuous light for 12 days and induced with 1% ethanol. After 12 hours of EtOH treatment, seedlings were dissected and RNA was processed from the shoot apex and from young leaves. Affymetrix ATH1 arrays were hybridized in duplicates for each gene construct and condition [LTB$^+$05]. In total we considered 14 different gene sets to be discriminated by the methods.

3.2 Experimental Setup

To train and test the method we first split the data into two parts (80%:20%). The first part is used for motif finding and SVM training. For hyper-parameter tuning we used the first part with 5-fold cross-validation to find the optimal combinations of hyper-parameters. (The SVM and the considered kernels have several hyper-parameters to be given in advance. This includes the regularization parameter C of the SVM, the maximal length of oligomers ℓ and the maximal shift S considered in the WDS kernel.) The second part is used for estimating the generalization performance. Here we measure the area under the ROC curve (auROC) as the generalization performance (random guessing corresponds to 50% auROC).

The above procedure is repeated five times for different splits of training and test examples (outer cross-validation loop). As performance measure we report the average auROC over the five splits.

[4] The fold change is computed from the normalized gene expression level p in treatment and respective control, c: $n = \begin{cases} -c/p & \text{if } p/c < 1 \\ p/c & \text{if } p/c \geq 1 \end{cases}$. In this case the direction of the change is represented by the sign of n, positive means up and negative means down relative to the control. If several replicates were available, the mean after normalization is taken for every gene, for all replicates of p and c respectively.

3.3 Results

In a first experiment we illustrate that simple methods as for instance SVMs with spectrum or WDS kernel cannot easily solve the considered classification problem. The results are given in Figure 2. We can observe that essentially for all gene sets SVMs with Spectrum kernel fails to identify positive genes (auROC close to 50%). The SVM with WDS kernel is slightly better, but still produces close to random predictions.

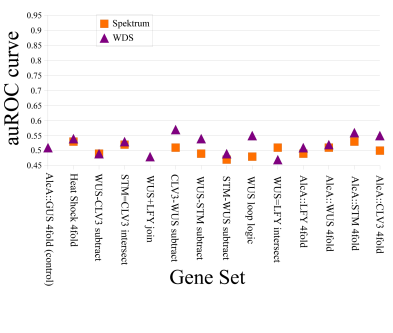

Figure 2: Accuracy of the Spectrum and WDS kernels: The prediction is rarely better than random guessing for these kernels. The kernels are not well suited for the this particular problem.

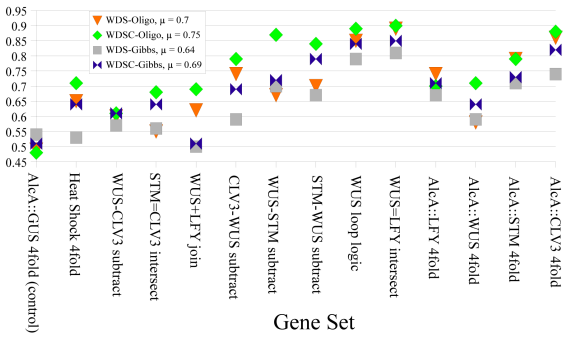

Figure 3: Accuracy of variations of the KIRMES approach: This graph shows a comparison of the basic kernels and the conservation kernels (C) combined with two different motif generation approaches: by oligo-counting (Oligo) or by Gibbs sampling (Gibbs). The average performance (μ) is given for each kernel variant. The first set is taken from a control experiment.

In Figure 3, we present results of the proposed methods in four variants: with motif discovery by Gibbs-sampling *vs.* oligo-counting as well as with and without using conservation (*cf.* Appendix A.2). We can make the following observations: (i) All four versions show a significantly improved performance relative to the base-line methods. (ii) Motif-finding using oligo-counting seems to work considerably better in combination with SVMs than Gibbs-sampling. A possible reason may be that the number of considered oligos (100-200) is higher than the number of motifs generated by the Gibbs-sampler (less than 50). (iii) Using conservation as weighting for the WDS kernel considerably improves the recognition performance. It results in an average improvement of 5 percentage points.

These results clearly illustrate the power of our approach in exploiting the relationship between motifs as well as the conservation to improve the recognition of transcription factor targets.

The algorithm can be used for any combination of regulatory regions and also any organism. Use of the web service integrated into Galaxy is straightforward and the resulting classification can help scientists with experimental microarray data select genes they want to investigate further.

An integration of protein binding data such as from chromatin immunoprecipitation experiments on a microarray chip is planned for a future extension of this method. Binding data can for example contribute to the weighting of a certain transcription factor binding

site and the surrounding sequence, just like conservation information. In that respect, the normalization scheme for the number of contributing related organisms can be remodeled to take into account their evolutionary distances and to generalize it further.

The use by experimentalists will ultimately determine the utility of this approach and govern the direction of further extensions together with technological advances such as Next Generation Sequencing methods for transcriptome or protein binding data.

Acknowledgments The authors thank the anonymous reviewers for their helpful suggestions that improved the manuscript. G.R. would like to thank Gabriele Schweikert for comments on the manuscript.

References

[Agi06] Agilent. GeneSpring GX. Technical report, Agilent Technologies, 2006.

[BE94] T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc. ISMB'94*, volume 2, pages 28–36, Menlo Park, California, USA, 1994. ISCB, AAAI Press.

[BGV92] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. COLT '92*, pages 144–152, Pittsburgh, Pennsylvania, United States, 1992. ACM Press.

[BOS+08] A. Benhur, C.S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. Support Vector Machines and Kernels for Computational Biology. *PLoS Computational Biology*, 2008. forthcoming.

[BWS05] W. Busch, M. Wunderlich, and F. Schoeffl. Identification of novel heat shock factor-dependent genes and biochemical pathways in A. thaliana. *Plant J*, 41(1):1–14, 2005.

[FSKB08] M.C. Frith, N.F. Saunders, B. Kobe, and T.L. Bailey. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol*, 4(4):e1000071, 2008.

[GL05] M. Gupta and J.S. Liu. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci U S A*, 102(20):7079–7084, 2005.

[GRH+05] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, and et al. Galaxy: a platform for large-scale genome analysis. *Genome Res*, 15(10):1451–1455, 2005.

[Joa99] T. Joachims. Making large-Scale SVM Learning Practical. In Bernhard Schölkopf, C Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA, 1999.

[LAB+93] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, October 1993. 0036-8075

[LEN02] C. Leslie, E. Eskin, and W. S. Noble. The Spectrum Kernel: A String Kernel For SVM Protein Classification. In *Proc. PSB'02*, pages 564–575, 2002.

[LEWN03] C. Leslie, E. Eskin, J. Weston, and W.S. Noble. Mismatch String Kernels for Discriminative Protein Classification. *Bioinformatics*, 20(4), 2003.

[LTB+05] A. Leibfried, J.P.C. To, W. Busch, S. Stehling, A. Kehle, M. Demar, J.J. Kieber, and J.U. Lohmann. WUSCHEL controls meristem function by direct regulation of cytokinin-inducible response regulators. *Nature*, 438:1172–1175, December 2005.

[MFG+03] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, and et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–8, January 2003. 1362-4962

[MMR+01] K.R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.*, 12:181–201, 2001.

[MTMM04] P. Meinicke, M. Tech, B. Morgenstern, and R. Merkl. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, 5(169), 2004.

[Nob06] W.S. Noble. What is a Support Vector Machine? *Nature Biotechnology*, 12(24):1565–1567, 2006.

[Pyt07] Python Software Foundation. Python. `http://python.org`, May 2007.

[RHTT04] J.C. Redman, B.J. Haas, G. Tanimoto, and C.D. Town. Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant J*, 38(3):545–561, 2004.

[RS04] G. Rätsch and S. Sonnenburg. Accurate Splice Site Detection for *Caenorhabditis elegans*. In K. Tsuda B. Schölkopf and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 277–298. MIT Press, 2004.

[RSS05] G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: Recognition of Alternatively Spliced Exons in *C. elegans*. *Bioinformatics*, 21(Suppl. 1):i369–i377, June 2005.

[SAE+04] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–4, January 2004. 1362-4962

[SRR07] S. Sonnenburg, G. Rätsch, and K. Rieck. Large Scale Learning with String Kernels. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*, pages 73–104. MIT Press, 2007.

[SRSS06] S. Sonnenburg, Gunnar Rätsch, C. Schäfer, and Bernhard Schölkopf. Large Scale Multiple Kernel Learning. *J of Mach Learn Res*, 7(Jul):1531–1565, July 2006.

[SS02] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[SS05] E. Segal and R. Sharan. A Discriminative Model for Identifying Spatial Cis-Regulatory Modules. *J of Comp Biol*, 12:822–834, 2005.

[SSGE86] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J Mol Biol*, 188:415–431, April 1986.

[SSP+07] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch. Accurate splice site prediction using SVMs. *BMC Bioinformatics*, 8(Suppl. 10):S7, 2007.

[ST02] Saurabh Sinha and Martin Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 30(24):5549–5560, 2002.

[Sto00] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, January 2000.

[STV04] B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel methods in computational biology*. MIT Press, Cambridge, MA, 2004.

[TMDS+02] G. Thijs, Y. Moreau, F. De Smet, J. Mathys, M. Lescot, S. Rombauts, P. Rouze, B. De Moor, and K. Marchal. INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, 18(2):331–2, February 2002.

[YTI+98] T Yada, Y Totoki, M Ishikawa, K Asai, and K Nakai. Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. *Bioinformatics*, 14(4):317–325, 1998.

[ZRM+00] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *Bioinformatics*, 16(9):799–807, September 2000.