PLOS ONE

RESEARCH ARTICLE

The folded *k*-spectrum kernel: A machine learning approach to detecting transcription factor binding sites with gapped nucleotide dependencies

Abdulkadir Elmas[1], Xiaodong Wang[1], Jacqueline M. Dresch[2]*

1 Department of Electrical Engineering, Columbia University, New York, NY, United States of America, **2** Department of Mathematics and Computer Science, Clark University, Worcester, MA, United States of America

* jdresch@clarku.edu

Citation: Elmas A, Wang X, Dresch JM (2017) The folded *k*-spectrum kernel: A machine learning approach to detecting transcription factor binding sites with gapped nucleotide dependencies. PLoS ONE 12(10): e0185570. https://doi.org/10.1371/journal.pone.0185570

Editor: Bin Liu, Harbin Institute of Technology Shenzhen Graduate School, CHINA

Received: July 11, 2017

Accepted: September 14, 2017

Published: October 5, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work has been supported by a National Institutes of Health (GM110571) grant (https://www.nih.gov) to JMD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Understanding the molecular machinery involved in transcriptional regulation is central to improving our knowledge of an organism's development, disease, and evolution. The building blocks of this complex molecular machinery are an organism's genomic DNA sequence and transcription factor proteins. Despite the vast amount of sequence data now available for many model organisms, predicting where transcription factors bind, often referred to as 'motif detection' is still incredibly challenging. In this study, we develop a novel bioinformatic approach to binding site prediction. We do this by extending pre-existing SVM approaches in an unbiased way to include all possible gapped *k*-mers, representing different combinations of complex nucleotide dependencies within binding sites. We show the advantages of this new approach when compared to existing SVM approaches, through a rigorous set of cross-validation experiments. We also demonstrate the effectiveness of our new approach by reporting on its improved performance on a set of 127 genomic regions known to regulate gene expression along the anterio-posterior axis in early *Drosophila* embryos.

Introduction

When studying the complex control of gene expression, often the first step is to locate an enhancer, or *cis*-regulatory module (CRM), within the genome. An enhancer is a non-coding region of DNA, typically located upstream of the promoter region, which binds transcription factor (TF) proteins and subsequently regulates the gene's expression. This regulation is extremely important for many of the key processes involved in embryonic development [1, 2]. The discovery of clusters of key TF binding sites has been critical to identifying potential enhancers in the genome, and mapping out the organization of these binding sites has been crucial in guiding our understanding of enhancer function and evolution [2–4].

DNA motif discovery and sequence classification to identify portions of the genome with specific biological function(s) has been a central problem in computational biology, addressed

by numerous approaches based on sequence alignment [5, 6], profiling consensus patterns of motifs [7, 8], and hidden Markov models [9, 10]. In this study, we draw from ideas introduced in Position Weight Matrix-based approaches to develop a novel, unbiased Support Vector Machine approach [11, 12].

Position Weight Matrix approach

Since the 1980s, one of the most popular bioinformatic approaches for predicting transcription factor binding sites (TFBSs) involves constructing a Position Weight Matrix (PWM) and scanning a DNA sequence for subsequences with a high probability of binding the TF of interest [13–15]. Standard implementations of this approach rely on the assumption that nucleotide positions within a TFBS are independent of each other [16, 17]. Even with this simplifying assumption, these models have proven themselves effective in predicting binding sites in a variety of different species [16–20].

In an attempt to implement a more systematic and unbiased approach in predicting TFBSs, there have been numerous different extensions to this approach, relaxing the assumption of nucleotide independence of contiguous nucleotides within the binding site. Common extensions include dinucleotide and *k*-mer models, which allow for dependence between adjacent nucleotides or a contiguous string of *k* nucleotides, respectively [21–23]. Even more recently, an algorithm was developed, referred to as MARZ, that combinatorially considers all possible gapped nucleotide dependencies across a fixed number of nucleotides [11]. By considering all possible dependencies, this new algorithm includes traditional PWM models, dinucleotide and *k*-mer models, as well as models with noncontiguous (i.e. gapped) dependencies. When tested on a set of well characterized TFs in *Drosophila*, MARZ illustrated that gapped models often outperform traditional PWM-based models [11, 12]. Although this may not be the case for all TFs or in all species, these studies have highlighted the importance of using an unbiased approach and considering all possible combinations of nucleotide dependence when attempting to make robust predictions of TFBSs.

Support Vector Machines

In recent years, discriminative approaches, such as Support Vector Machines (SVM), have been introduced and shown to be the best performing methods for sequence classification [24]. In the SVM classifier approach, the input sequences from different classes are considered as labeled examples and a learning algorithm is trained to find an optimal decision boundary between the different classes. The decision boundary (determined by a hyperplane in a multi-dimensional feature space) optimally separates the feature representations of the labeled examples. In this supervised approach, the unlabeled (test) sequences are later given to the algorithm, and the algorithm uses the learned decision boundary to predict labels/classes for these test sequences [25].

One of the earliest SVM approaches for biological sequence classification, which was originally implemented on protein sequences made up of strings of amino acids, was the Fisher kernel [26]. This approach, tested on the SCOP database [27], is based on a computationally-demanding generative model that requires one to build hidden Markov model profiles for each positive training sequence to obtain the feature vector representations. A new protein sequence is represented by a Fisher score vector, then its Kernel function for a given protein family is computed based on the Euclidean distance between this score vector and the score vectors that are precomputed for the known positive and negative examples of that protein family [26]. In a later study by [28], the SVM-pairwise kernel was introduced in which the pairwise alignments between each training sequence are used as the SVM features. This

methodology is similar to the Fisher kernel approach and the process is computationally-expensive as well. In [29], a more efficient "spectrum" kernel was introduced for the classification of protein sequences, where the feature vector consists of the occurrences of all *k*-mer contiguous subsequences in a given amino acid sequence. In subsequent works, the authors extended their approach for "inexact" subsequence matches for the improved classification performance, which allows up to a certain number of mismatches when counting the *k*-mer occurrences [30, 31].

The advantage / extension of our approach. The *k*-mer spectrum is an effective representation of DNA sequences in terms of discriminating functional segments (i.e., TFBS, or promoter/enhancer regions) from the genomic background. The TFBSs can be characterized by the specific composition of the *k*-mer subsequences inherent to the binding preference of a TF protein, where the consensus motif sequence is defined as the most observed *k*-mer in the corresponding binding sites followed by those that differ from the consensus sequence in various degrees.

It has been suggested that protein binding sites have varying levels of nucleotide interdependencies, i.e., the nucleotide patterns in certain (non-adjacent) bases within binding sites may appear more often than the patterns in other bases [12, 32, 33]. This suggests a (gapped) dependency in some non-adjacent bases within the TFBSs which can be modeled by a "gapped" *k*-mer. The gapped *k*-mer is a length *k* sequence of nucleotides that breaks dependency in certain bases that cannot form a significant consensus, i.e., we can assume the interdependency between the first and the last nucleotides in the following 3 binding sites {"*AACA*", "*ACGA*", "*AGTA*"}, where a "gapped" *k*-mer represented by "*ANNA*" (where each '*N*' represents a gap, corresponding to no specific nucleotide preference) will be a better fit to the majority of TFBSs (i.e., 3/3) than any other "contiguous" *k*-mer.

Following this intuition, to account for the TF proteins that possess nucleotide interdependency, a more suitable representation of the DNA binding sites –in terms of sequence discrimination– should incorporate the gapped *k*-mer compositions as well as the regular *k*-mers. In that respect, the composition (enrichment) of the specific gapped *k*-mers in the analyzed sequence (relative to the enrichments obtained from the training data) can help identify such TFs that possess interdependency in their preferred binding sites. In the following section, we describe a generalized discriminative approach for detecting functional DNA sequences including those that may possess nucleotide interdependency in particular TFBSs.

Although PWM-based approaches are often used for TFBS discovery, and have shown success in incorporating gapped dependencies, SVM approaches require much less prior knowledge of binding preferences [11, 12, 24, 25]. A PWM is built from a set of sequences known to bind a specific TF [11, 16, 17]. Thus, these studies focus on specific sets of TFs, often with the goal of identifying high affinity binding sites, and have no flexibility in discovering TFBSs for TFs not included in the initial input [11, 16, 17]. SVM approaches, on the other hand, require no knowledge of the TFs that may bind to the DNA search sequence or their binding preference within the genome; they are searching for enriched motifs within the search sequence with the goal of identifying TFBSs [24, 25]. Due to the extremely different inputs needed, and often-different underlying goals of using such algorithms, it is impossible to conduct a valid comparison of the results from PWM-based vs. SVM approaches.

SVM for sequence classification

Support Vector Machine (SVM) is a discriminative learning method that finds an optimal decision boundary that separates, in the case of binary classification, the positive and negative data sets represented in a high-dimensional vector space [25]. Consider the training data set of

labeled input vectors $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, t$, where $\boldsymbol{x}_i \in \mathbb{R}^N$ is the projection of the i-th input data into a high-dimensional feature space (obtained by a known mapping function, Ψ), and $y_i \in \{-1, 1\}$ is the respective class label. In this classification problem, a basic (linear) decision boundary $\mathcal{B}$ can be found by minimizing $||\boldsymbol{w}||^2$ subject to $y_i(\boldsymbol{x}_i^T\boldsymbol{w} + b) \geq 1$, $i = 1, \ldots, t$, where $\boldsymbol{w} \in \mathbb{R}^N$ is the normal vector to the decision boundary (which is a hyperplane, i.e., $\mathcal{B} = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}^T\boldsymbol{w} + b = 0\}$), b is the classifier bias, and the term $\boldsymbol{x}_i^T\boldsymbol{w}$ represents the inner product between the vectors $\boldsymbol{x}_i$ and $\boldsymbol{w}$. For practical considerations a dual problem is rather solved to find the decision boundary by forming the Lagrangian (with multipliers α) for the above quadratic programming problem [24]:

$$\begin{aligned} &\max \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j, \\ &\text{subject to} \quad \alpha_i \geq 0 \forall i, \end{aligned} \tag{1}$$

whereby the normal vector $\boldsymbol{w}$ can be optimally constructed from $\boldsymbol{w} = \sum_i y_i\alpha_i\boldsymbol{x}_i$. Here, the learned variables ($\boldsymbol{w}$, b) serve as the parameters of the classification rule. A test example $\boldsymbol{z}$ is then classified by the expression

$$f(\boldsymbol{z}) = \boldsymbol{z}^T\boldsymbol{w} + b = \sum_i y_i\alpha_i\boldsymbol{z}^T\boldsymbol{x}_i + b,$$

where $f(\boldsymbol{z})$ represents the distance of $\boldsymbol{z}$ from the learned decision boundary, and its sign gives the estimated class label. The inner product in (1) yields a measure of similarity in the feature space $\mathcal{F}$ where the coordinates are defined by the vector elements. The similarity between any two input data (χ_i, χ_j) can be generalized by using the Kernel functions $K(\chi_i, \chi_j)$ [24].

A convenient Kernel function for sequence classification is the *k-spectrum kernel*, [29] which describes the similarity of sequences by their (k-mer) subsequence compositions of a fixed length k. Consider that the training data χ is a character sequence belonging to an input space $\mathcal{X}$ consisting of all finite sequences from an alphabet $\mathcal{A}$ of size ℓ. For a given $k \geq 1$, it is projected to a frequency vector called "*k-spectrum*", i.e., $\boldsymbol{x} = \Psi_k(\chi)$, that represents the occurrences of all possible k-mer subsequences in χ

$$\Psi_k(\chi) = (\phi_{\boldsymbol{\alpha}}(\chi))_{\boldsymbol{\alpha}\in\mathcal{A}^k},$$

where $\phi_{\boldsymbol{\alpha}}(\chi)$ is the number of times $\boldsymbol{\alpha}$ occurs in χ, and $\mathcal{A}^k$ represents the set of all possible length-k sequences from the alphabet $\mathcal{A}$.

In [29], the k-spectrum kernel's mapping function Ψ_k considers only the subsequences of "contiguous" (dependent) nucleotides (i.e., k-mers). In a later related work, authors introduce the *mismatch kernel* to allow for a certain number of mismatches in the k-mer occurrences. Recently, a more systematic approach was proposed for detecting regulatory sequences possessing certain degrees of nucleotide interdependency through the search of sequence motifs called "gapped k-mers" [11]. In our study, we extend the concept of k-spectrum given in [29] by incorporating all types of nucleotide interdependency in k-mers to improve the predictive power of the feature set. That is, we extend the feature set of k-mers by those of the gapped k-mers given in [11] that ignore all possible subsets of nucleotides. As it will be explained next, the mapping function only evaluates the data points under the k-spectrum and computes the extended (gapped) features efficiently by folding this spectrum in certain coordinates specific to the gapped k-mer features.

Gapped k-mers have been recently used for regulatory sequence prediction in the SVM-based classifiers. In [34], the authors consider ℓ-mers with k non-gapped bases ($k \leq \ell$) and estimate ℓk feature frequencies based on the ℓ-mer mismatch profiles between sequences. Their

model considers a fixed number of gaps, i.e., $\ell - k$, when computing the SVM kernel. In contrast, we allow "variable-length" gaps (between 0 and $k - 1$) in the *k*-mer and thereby incorporate a full and unbiased list of sequence features in the SVM kernel. This also allows us to perform systematic analyses on the sequences which we describe next.

In the context of gapped dependency, another recent study [35] employed an SVM learning approach, repDNA, in which pseudo nucleotide compositions that represent the gapped dependency of nucleotides is incorporated into the *k*-mers, as well as more sophisticated features derived from the physicochemical properties of DNA.

Folded *k*-spectrum kernel

We use the binary notation to refer the dependent nucleotides in a subseqeunce $\boldsymbol{\alpha}$, where '1' represents the dependent nucleotides and '0' represents the gap. For example, given $k = 3$ (3-mer) the binary value of the decimal number 5 is 101; therefore 5 encodes the 3-mer sequences of which only the 1st and the 3rd nucleotides are interdependent. We denote this dependency model by a set of gapped *k*-mer features $\mathcal{A}_5^3$ which consists of 16 sequences varying on the nucleotides (1,3), i.e., $\mathcal{A}_5^3 = \{ANA, ANC, ANG, ANT, CNA, \ldots, TNG, TNT\}$, where the gaps are represented by 'N'. Similarly, the decimal number 7 yields the binary sequence 111 and the corresponding $\mathcal{A}_7^3$ represents the set of all (contiguous) 3-mer sequences. For $k = 3$, the remaining possible feature sets are $\mathcal{A}_3^3$ and $\mathcal{A}_1^3$ which consists of all possible dimer (011) and monomer (001) sequences, respectively.

As described above, there are 4 different feature sets (gapped *k*-mer models) for $k = 3$, i.e., monomer, dimer, 1-gapped, and 3-mer. In general, this number grows exponentially with *k*, i.e., 2^{k-1} different feature sets. However, in practice the length of $k = 6$ is a sufficient (and optimal) choice for the purpose of sequence discrimination [36, 37], which we also used in this study. In the following, we present a general formulation of the folded *k*-spectrum kernel for any *k*.

Using the same notation, we define the feature map of a particular feature set $\mathcal{A}_{2m-1}^k, m = 1, \ldots, 2^{k-1}$ by

$$\Phi_{(k,m)}(\chi) = (\phi_{\boldsymbol{\alpha}}(\chi))_{\boldsymbol{\alpha} \in \mathcal{A}_{2m-1}^k}, \tag{2}$$

where $\Phi_{(k,m)}$ maps the input data into the *m*-th feature set which can be denoted by $\mathcal{A}_{2m-1}^k = ([\alpha\alpha \ldots \alpha] \otimes (2m-1)_{(2)})_{\alpha \in \mathcal{A}}$, with $(l)_{(2)}$ representing the decimal *l* in binary sequence, i.e. $(5)_{(2)} = 101$. In Eq (2), we used the relative frequency mapping for $\phi_{\boldsymbol{\alpha}}(\chi)$ to cancel out the influence of different sequence lengths and different background frequencies of $\boldsymbol{\alpha}$, i.e., it represents the number of times $\boldsymbol{\alpha}$ occurs in χ divided by the number of possible *k*-mers ($|\chi| - k + 1$), relative to the same measure observed in the genome

$$\phi_{\boldsymbol{\alpha}}(\chi) = \frac{\#\,\boldsymbol{\alpha}\ occurs\ in\ \chi}{|\chi| - k + 1} \left(\frac{\#\,\boldsymbol{\alpha}\ occurs\ in\ genome}{|genome| - k + 1} \right)^{-1}.$$

Notice that in (2), a value of $m = 2^{k-1}$ corresponds to the last feature set (which is the set of contiguous *k*-mers), where the term $2m - 1$ yields the length-*k* sequence of 1s in the binary representation, i.e., $(2m-1)_{(2)} = (2 \times 2^{k-1} - 1)_{(2)} = (2^k - 1)_{(2)} = \{1\}^k = [11\ldots1]$. Similary, $m = 1$ represents the monomer model with the feature set

$$\mathcal{A}_{2m-1}^k = \mathcal{A}_1^k = \{N \ldots NA,\ N \ldots NC,\ N \ldots NG,\ N \ldots NT\},$$

and $m = 2$ represents the dimer model with the feature set

$$\mathcal{A}^k_{2m-1} = \mathcal{A}^k_3 = \{N\ldots NAA,\ N\ldots NAC,\ N\ldots NAG,\ N\ldots NAT,\ N\ldots NCA,\ \ldots,\ N\ldots NTT\}.$$

Any other value between $2 < m < 2^{k-1}$ corresponds to the gapped feature sets given that $k > 2$. Considering all possible gapped *k*-mers, i.e., $\{m = 1, \ldots, 2^{k-1}\}$, the extended *k*-spectrum feature map is then given by

$$\Psi_k(\chi) = (\Phi_{(k,m)}(\chi))_{m=1}^{2^{k-1}}. \tag{3}$$

For efficiency, we only evaluate the contiguous *k*-mer features under the *k*-spectrum map, i.e., $\Phi_{(k,2^{k-1})}(\chi)$, then the mapping of χ to any other feature set ($m < 2^{k-1}$) is calculated as a linear sum of the feature-specific values in this *k*-spectrum map, imitating a folding process over those coordinates. For example, given $k = 3$ the frequency of the feature sequence "*ANA*" in $\mathcal{A}^3_5$, is computed by

$$\phi_{ANA}(\chi) = \sum_{\alpha\in\mathcal{A}} \phi_{A\alpha A}(\chi) = \phi_{AAA}(\chi) + \phi_{ACA}(\chi) + \phi_{AGA}(\chi) + \phi_{ATA}(\chi),$$

where $[\phi_{AAA}(\chi), \phi_{ACA}(\chi), \phi_{AGA}(\chi), \phi_{ATA}(\chi)]$ are obtained from $\Phi_{(k,2^{k-1})}(\chi)$. In other words, the mapping function of a gapped *k*-mer feature (i.e., "*ANA*") is the (unweighted) linear combination of the mapping functions of all contiguous *k*-mer features which differ from that gapped *k*-mer sequence in the gapped locations.

Additional approaches

Many alternative approaches have been recently proposed in motif discovery and feature prediction. One such approach investigates the gapped nucleotide dependencies in terms of discriminating real pre-miRNA sequences from false sequences. In [38], authors use *k*-mer components to represent the RNA sequences, and they propose an approach (degenerate *k*-mer) to deal with the problem of over fitting. In another study, this method was modified and applied to DNA-binding proteins [35] based on building pseudo amino acid composition profiles and then incorporating gapped dependency between amino acid pairs through a reduced alphabet of amino acids. Features such as these pseudo components of DNA, RNA, or protein sequences are generated using previously published and available tools such as repDNA and Pse-in-One. [39, 40]. The approach was evaluated by the Support Vector Machines classification.

Another computational method was proposed in [41], referred to as WSMD, in which the optimal set of motifs and the sequences containing them are simultaneously identified through a weakly supervised learning method. Although the method does not incorporate the gapped dependency in the identified motifs, using a similar approach, referred to as LMMO, which maximized the classification accuracy and other relevant metrics (AUC [42]), the algorithm was able to discover a variety of regulatory motifs using a continuous global optimization scheme.

Materials and methods

Data sets

We tested the proposed SVM kernel through the binary classification problem, i.e., discriminating a single class of positive (functional) DNA sequences from negative (background) DNA

sequences. As the positive signal set, we used the set of 127 CRMs studied in [43], belonging to 114 genes of interest in early *Drosophila* development that exhibit differential expression patterns along the anterio-posterior (AP) axis. The CRM sequences were originally obtained from the data base [44] and extended by 100-bps of flanking sequence in each direction. We generate the negative signal set by retaining the same distribution of sequence lengths and nucleotide compositions as the positive signal set. That is, for each sequence in the positive set, we generate 10 different scrambled versions using random (uniformly-selected) permutations of nucleotides from the original sequence.

Although using a larger negative set generally leads to the problem of an imbalanced data set, we have not observed any imbalance in our data set; as shown in our results, the proposed approach performs with excellent precision-recall curves. If our algorithm resulted in suboptimal performance due to class imbalance on a particular data set, conventional approaches could be applied from the fields of machine learning and bioinformatics, such as data sampling [45] [46], sample weighting [47] [48] [49], cost-sensitive learning [50] [51], kernel-based methods [52], or hybrid approaches [53].

All positive and negative data sets used in this study are available from http://www.columbia.edu/∼ae2321/DmelCRMs.zip.

Identifying top-enriched features

Through SVM training, one can investigate the enrichment of the individual features (gapped *k*-mers) observed in a given input sequence. For this, we use all positive and negative data sets and train the SVM with the proposed kernel. The elements $w(n)$, $n = 1, \ldots, N$ in the resulting decision (weight) vector $\boldsymbol{w} = [w(1), \ldots, w(N)] \in \mathbb{R}^N$ may indicate the relative discriminative power of the corresponding features, whereby the class of a test example mapped into this feature space $\boldsymbol{x} = [x(1), \ldots, x(N)] \in \mathbb{R}^N$ is obtained through the sum of the weighted frequencies of its sequence features, i.e. the sign of the inner product $\boldsymbol{w}^T\boldsymbol{x} = \sum_{n=1}^{N} w(n)x(n)$. Following this intuition, given the test sequence $\boldsymbol{x}$ we define the enrichment score of the n-th feature in this sequence by $r(n) = w(n)x(n)$ and the score vector corresponding to all features by $\boldsymbol{r} = [r(1), \ldots, r(N)] \in \mathbb{R}^N$.

Considering the DNA sequence data, different subsets of features may be enriched in different sequences due to the underlying degeneracy of the binding sequence [54]. In addition, since the nature of DNA binding prefers a "consensus" for particular nucleotides rather than the exact binding sequence, *k*-mers containing a preferred core sequence with certain mismatches may also be enriched [31]. Thereby, it is intuitive to expect that a number of top-enriched features estimated for a given sequence may represent binding sites for the underlying regulatory factor(s) driving the CRM's primary function [36].

To investigate this, we calculate the enrichment vectors for every positive sequence used in the SVM training (i.e., CRM data set), i.e., $\{\boldsymbol{r}_i, i = 1, \ldots, P\}$. For each given sequence, we select the top-enriched features residing above a cutoff weight (0.005) and those belonging to the same feature set (gapped *k*-mer model) are used to construct the relevant "gapped *k*-mer motif". For example, suppose that in the top-enrichment results there are only two features *ANT* and *TNT* belonging to the set $\mathcal{A}_5^3$, then the corresponding gapped *k*-mer motif of $\mathcal{A}_5^3$ will be *(A/T)NT*. In fact, such a motif may be the fragment of a real DNA binding motif (i.e., transcription factor binding site), whereby using this *k*-mer motif one can search for the possible hits in a known motif data base. By this motivation, we generated all such motif fragments estimated for each given sequence in the CRM data set.

Filtering out potential false positive gapped *k*-mers (false discovery-based feature elimination)

It is known that certain sequence patterns may be found more frequently in the genomic background, entailing the danger, for the above procedure, of falsely discovering such patterns (features) and predicting the subsequent (false) motif hits [55–57]. To address any false discovery caused by the underlying nucleotide composition, we set out to eliminate the features that putatively lead to false enrichments. This is done by replacing the positive inputs with a set of "false" positive sequences to identify an ensemble of corresponding falsely-enriched features (Fig 1). That is, we replace each positive training sequence $\boldsymbol{\chi}_i$, $i = 1, \ldots, t$ with 10 scrambled versions of it ($\boldsymbol{\chi}_j^f, j = i, \ldots, i+9$), and use this expanded (false) positive set for training the SVM with the proposed kernel; then we estimate the top-enriched gapped *k*-mers (with $r_j^f(n), n = 1, \ldots, N$) for each of the false positive sequences $\boldsymbol{\chi}_j^f, j = 1, \ldots, 10t$, and discard those residing below the cutoff weight (0.005), i.e., set $r_j^f(n) = 0$ if $r_j^f(n) < 0.005$.

One should note that the 127 CRMs in the positive set are not an exhaustive list of CRMs driving expression of genes in early *Drosophila* development that exhibit differential expression patterns along the AP axis [43, 44]. Therefore, to obtain "false" positive sequences, we randomly scramble the positive sequences, retaining the same number of A, C, G and T nucleotides, opposed to a strategy of choosing random sequences from the genome [11, 12, 58].

Comparing top-enriched features with those in JASPAR

After filtering, we compared the individual gapped *k*-mers against a relevant motif data base (i.e., JASPAR's core data base for insects) [59] to elucidate the putative regulatory factors. For this task, we used the motif comparison tool Tomtom [60] which calculates statistical measures (p-value) to quantify the similarity between two motifs. For each sequence, we obtained all the motif hits found by Tomtom and retained those with a significant p-value score $< 10^{-3}$.

To improve the filtering procedure, we further discard "insignificant" false positives by filtering out any gapped *k*-mer which does not lead to a significant JASPAR hit. That is, for each false positive input $\boldsymbol{\chi}_j^f, j = 1, \ldots, 10t$, we construct all enriched gapped *k*-mer motifs and search the motif database via Tomtom. If a gapped *k*-mer motif results in a significant hit to a JASPAR motif (p-value < 0.001), then we keep this gapped *k*-mer motif, assuming that it consistently appears on the genomic background; otherwise if the motif is not found in JASPAR, we discard that motif, regarding it as an "insignificant" false positive.

The remaining gapped *k*-mers (features) are referred to as "false" and are used to remove the corresponding features found in the positive training sequence's top-features. That is, we filter out the n-th gapped *k*-mer (set $r_i(n) = 0$) if it belongs to one of the "gapped models" of the falsely-enriched features (corresponding to $r_j^f(n)$) in the respective scrambled copies $j = i, \ldots, i + 9$. After removal, the remaining gapped *k*-mers (with enrichment scores $r_i^h(n) > 0.005$) at each input sequence are assumed to be "high-confidence" predictions, whereby we constrained the feature set to this collection, i.e., $\boldsymbol{\Psi}_k^h = (\cup_i \mathcal{I}(\boldsymbol{r}_i^h)) \cap \boldsymbol{\Psi}_k$ where $\mathcal{I}(\boldsymbol{r}_i^h)$ represents the indices of the remaining features with $r_i^h(n) > 0.005, n \in \{1, \ldots, N\}$ (Fig 1).

Cross-validation

We assess the performance of the proposed folded *k*-spectrum kernel through cross-validation tests. During this procedure, a true positive was defined as a CRM from the original set of 127 'positive' CRMs that was not used in the initial training step but was detected by the algorithm (i.e., correctly classified as a positive sequence). Similarly, a false negative was defined as a

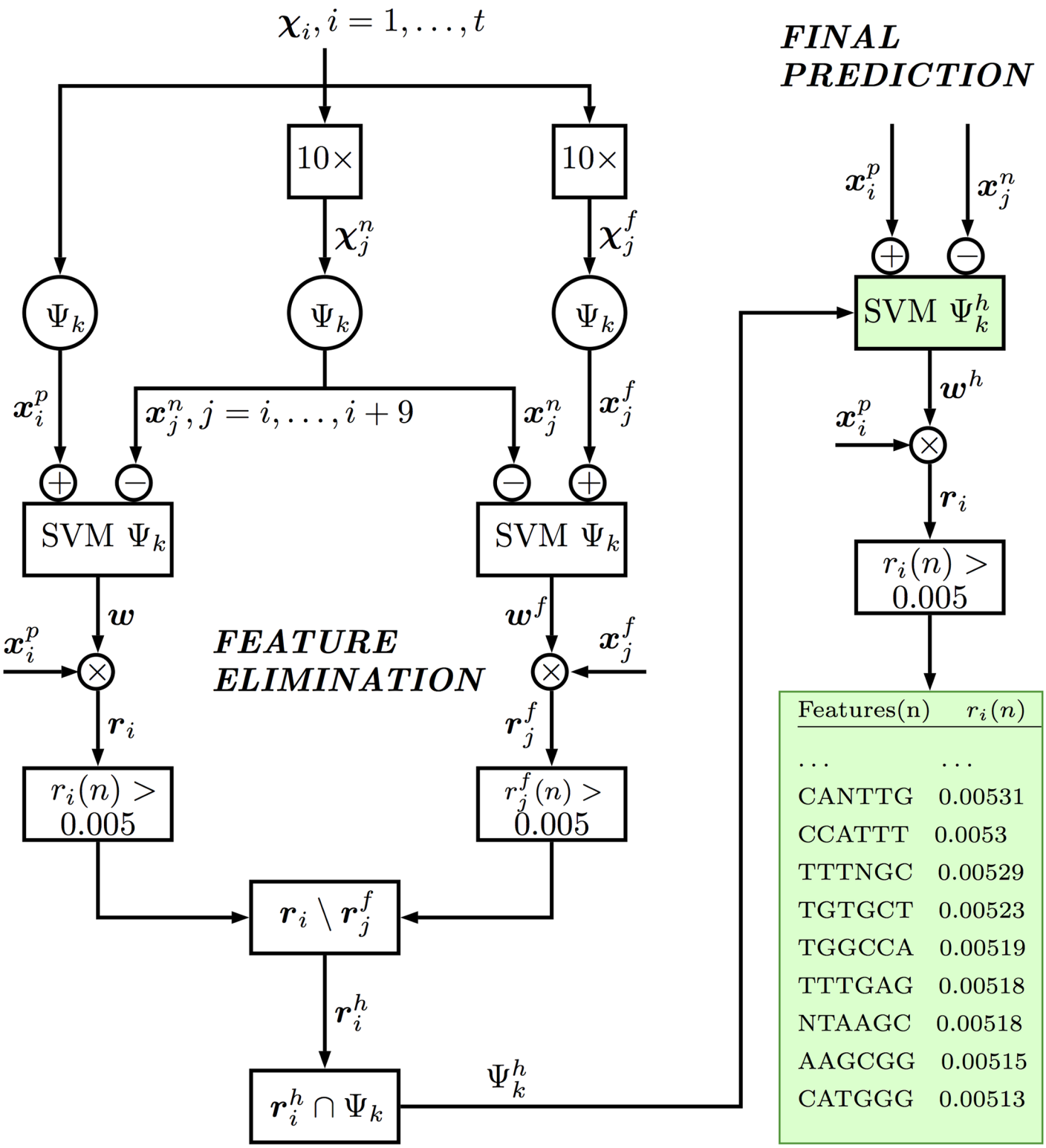

Fig 1. SVM with feature elimination based on the discovery of false enrichments. Feature enrichments (***r***) are estimated by SVM learning and a cutoff threshold. Then the false enrichments are detected and eliminated through using the background (scrambled) positive sequences (x_j^f) in the SVM learning. After elimination, the remaining features (r_i^h) are used in the final SVM model Ψ_k^h.

https://doi.org/10.1371/journal.pone.0185570.g001

CRM from the original set of 127 'positive' CRMs that was not used in the initial training step and was not detected by the algorithm (i.e., incorrectly classified as a negative sequence). Thus, any CRM with a low number of false negatives after this analysis represents a CRM that, even when left out of the training data, can still be detected as a positive sequence.

Since our data set consists of experimentally verified CRMs, we operate under the assumption that the set contains a large number of TFBSs [44]. Therefore, for a valid assessment of performance, we implement a 10-fold cross validation (i.e., we train the algorithm on a large majority of the inputs and test it on the remaining subset).

Given a set of input sequences (i.e., the CRM data set) with positive and negative class labels we randomly divide them into 10 disjoint subsets by retaining (approximately) equal proportions of the positive/negative sequences. At each fold, one different subset is left for testing and the other 9 subsets are used for training the SVM with the proposed folded *k*-spectrum kernel. We note that the false positive features are eliminated from the feature set and the feature map is constrained to Ψ_k^h following the procedure described in the previous sections.

After processing 10 folds, each input sequence is tested (classified) once. We then obtain those class predictions and determine the true/false predictions for the positive and negative sequences. The corresponding ROC curve and Area Under Curve (*AUC*) score are then generated (Fig 2).

We re-evaluate these predictions in the repeated experiments in order to average out the influence of randomization on the test subsets. We repeat the above procedure 100 times by randomly dividing the data in each experiment, then check the cumulative "false negative" prediction of each input sequence, which represents the number of times a trained classifier fails to detect that sequence accurately. This evaluation elucidates the predictive power of the "gapped *k*-mer" features on the individual input sequences.

We want to identify the group of CRMs in which the classifier performs with low accuracy, i.e., through repeated experiments the classifier "consistently" fails to detect those CRMs and results in a high number of false negative (FN) calls. The consistency can be defined by setting a threshold in the FN call rate, i.e., FN>10/100. For example, if a classifier (falsely) predicts a positive test sequence as negative in at most 10% of the repeated (randomized) cross-validation experiments, we can assume that the classifier is able to predict that sequence with high (90%) accuracy. Fig 2 displays the (sequence-specific) prediction accuracy of the classifiers, the contiguous *k*-spectrum kernel and the folded *k*-spectrum kernel, based on the 10% FN call rate cutoff. Subsequently, one can determine the groups of CRMs that the methods (contiguous *k*-mer vs. folded *k*-mer) perform differently or similarly, and investigate the underlying result of each classifier on these CRM groups.

Results

Identification of enriched motifs

We implemented the folded *k*-spectrum kernel approach using a set of 127 CRMs responsible for the regulation of 114 genes exhibiting differential expression during early *Drosophila* development. These CRMs are referred to as our 'positive sequences', as each of them has been experimentally shown to regulate gene expression along the AP axis, and thus each is thought to contain motif(s) corresponding to TFBSs present in the early embryo [61]. Those referred to as 'negative sequences' during our SVM training procedure have been constructed by randomly scrambling the positive sequences while retaining each sequence's original nucleotide distribution (see Methods for more details).

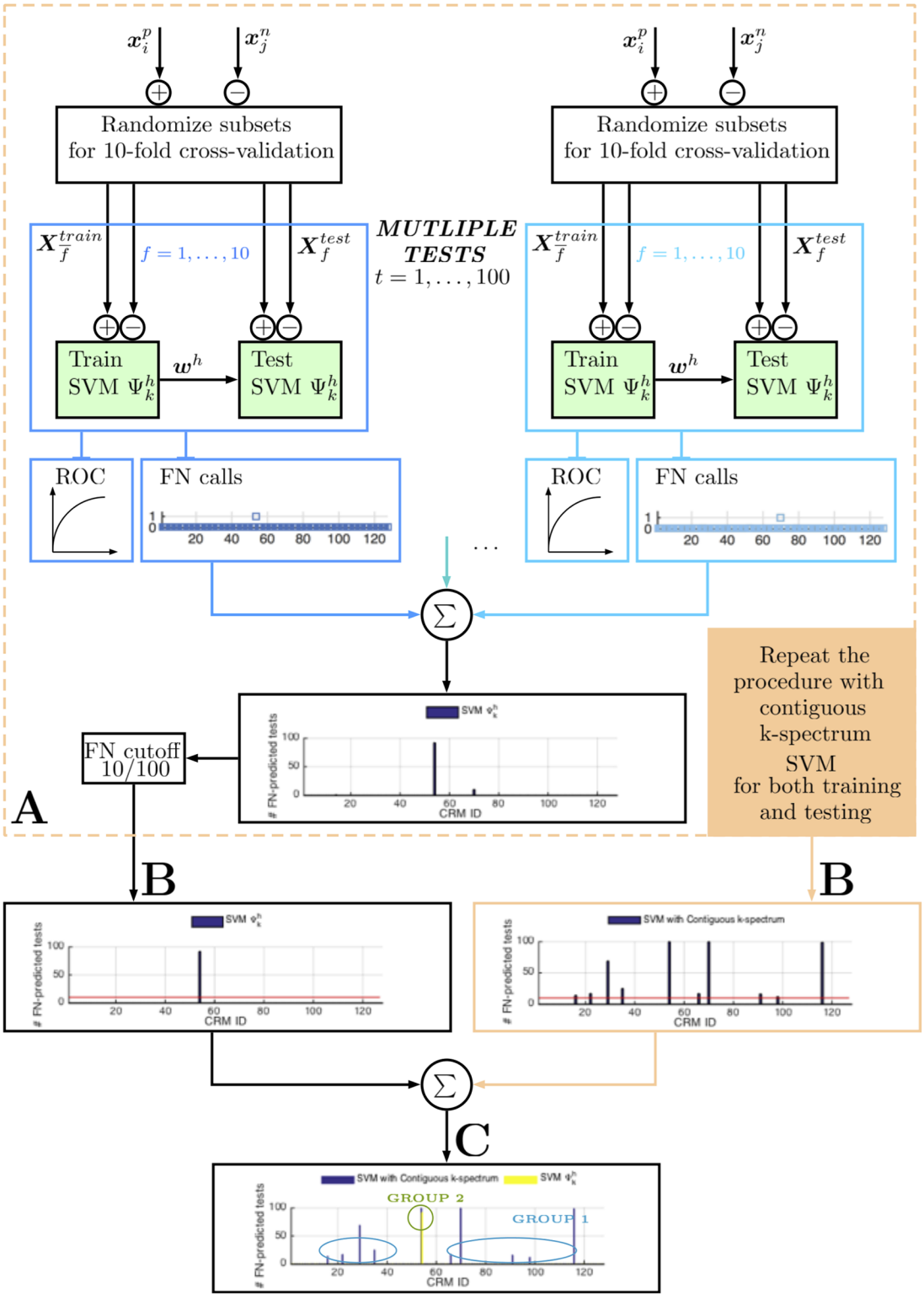

Fig 2. Discovery of the different groups of sequences in the two methods' (contiguous *k*-mer SVM, folded *k*-mer SVM) predictions, based on the differential FN (false negative) calls cumulated over multiple cross-validation tests. For each method, multiple randomized leave-sets-out cross-validation tests are performed and an ensemble of FN-predicted sequences are determined (**A**). The sequences that are called negative above 10% call rate (FN cutoff) are then reported (**B**). The reported sequences of both methods are compared and the intersecting/differing groups of sequences are determined (**C**).

https://doi.org/10.1371/journal.pone.0185570.g002

Based on the cross-validation tests described in Fig 2 (with 10% FN call rate cutoff), our algorithm resulted in the identification of three very interesting subsets (groups) of these CRMs:

Group 1. those that are correctly detected as belonging to the 'positive sequence set' by gapped *k*-mers (i.e., folded *k*-spectrum kernel), but not detected by contiguous *k*-mers (i.e., contiguous *k*-spectrum kernel),

Group 2. those that are incorrectly identified as belonging to the 'negative sequence set' by both gapped *k*-mers and contiguous *k*-mers, and

Group 3. those that are correctly detected as belonging to the 'positive sequence set' by both contiguous *k*-mers and gapped *k*-mers.

We observed that 7 CRMs fall into Group 1, 7 fall into Group 2, and the remaining 113 fall into Group 3 (Fig 3A, S1 Table). One should note that none of the 127 CRMs were correctly detected by contiguous *k*-mers but not by gapped *k*-mers due to the fact that contiguous *k*-mers are included when considering all possible gapped *k*-mers. The subset of CRMs contained in Group 1 may contain crucially important TFBSs with gapped motifs, which are undetectable using the standard approaches. Therefore, we focus our attention on those motifs

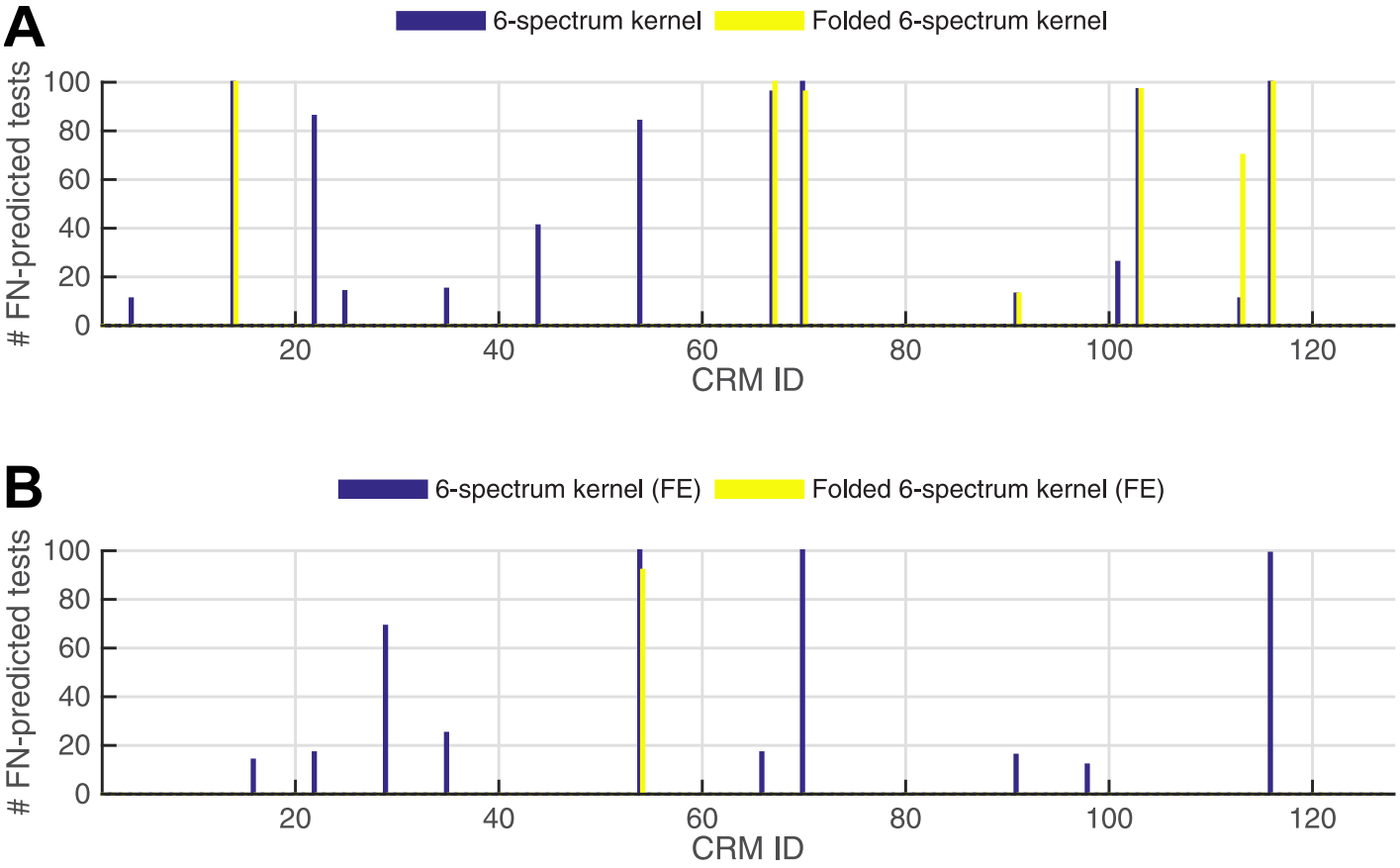

Fig 3. Classification performance of the (contiguous) *k*-spectrum kernel vs. folded *k*-spectrum kernel without (A) and with (B) feature elimination (FE) procedure. For each method, 100 randomized leave-sets-out cross-validation tests are performed and an ensemble of FN-predicted sequences are determined. The bar plot displays the total number of times a given (positive) CRM sequence is classified as negative.

https://doi.org/10.1371/journal.pone.0185570.g003

(gapped *k*-mers) enriched in the Group 1 CRMs (S2 Table), and include the enriched gapped *k*-mers in the Group 2 and 3 CRMs in the Supplementary Information (S3 Table).

The results in S2 Table give a comprehensive picture of all of the gapped *k*-mers found to be enriched (weight > 0.005) in the Group 1 CRMs. However, to further investigate the importance of particular gapped *k*-mer models, we combined these results into S4 Table, which gives each gapped *k*-mer model found to be enriched, the number of enriched gapped *k*-mers that fell into that gapped *k*-mer model, and the cumulative weight of those enriched gapped *k*-mers. Note that the most highly enriched gapped *k*-mers in each of the Group 1 CRMs are those containing only a small number of gaps (i.e. most have a single nucleotide gap), suggesting that these gaps are contributing significantly to TFBS identification.

Removing false positives

There are some sequence features that may be found to be more abundant than others, even in the genomic background sequence, due to the sequence composition, although they do not represent TFBSs. Thus, in an attempt to remove these 'false positive' motifs (spurious gapped *k*-mers), we have used a filtering procedure (see Methods for more details). This procedure has eliminated approximately 88% of the total gapped *k*-mer features that were previously found to be enriched. S5 and S6 Tables, and all remaining figures correspond to these filtered results.

After filtering, 9 CRMs fall into Group 1, 1 falls into Group 2, and the remaining set of 117 CRMs fall into Group 3 (Fig 3B, S7 Table). Note that 3 CRMs moved from Group 2 to Group 1 after filtering (i.e., the filtering procedure allowed them to be correctly detected as belonging to the 'positive sequence set' by gapped *k*-mers, but still incorrectly identified them by contiguous *k*-mers), only 1 CRM moved from Group 1 to Group 2 after filtering (i.e., the filtering procedure caused the gapped *k*-mers to lose their ability to correctly identify the CRM). Thus, overall the filtering procedure improves the performance of the gapped *k*-mers when compared to the contiguous *k*-mers.

S5 Table shows the gapped *k*-mers found to be enriched in the Group 1 CRMs using the 127 positive CRM sequences for training after eliminating features characterized as 'false' during this procedure.

The overall number of enriched gapped *k*-mers in the Group 1 CRMs decreases after filtering (from an average of approximately 176 enriched gapped *k*-mers per CRM to 90 enriched gapped *k*-mers per CRMs, S2 vs. S5 Tables), but there still remain a large number of enriched gapped *k*-mers.

Again, to investigate the importance of particular gapped *k*-mer models, we combined the filtered results into S6 Table, the number of enriched gapped *k*-mers within that model, and the cumulative weight of those enriched gapped *k*-mers. An interesting observation is that those found to be highly enriched after filtering have a small number of gaps, i.e., $\alpha\alpha\alpha N\alpha\alpha$, $\alpha N\alpha\alpha\alpha\alpha$, $\alpha\alpha\alpha\alpha N\alpha$, $\alpha\alpha N\alpha\alpha\alpha$ etc., where $\alpha \in \{A, C, G, T\}$.

However, one should note that the gapped *k*-mers that belong to the contiguous model $\alpha\alpha\alpha\alpha\alpha\alpha$ were characterized as false enrichments in all cases.

Improved performance of the folded *k*-spectrum approach

To validate our model and compare it to the traditional *k*-spectrum kernel, we conducted a set of cross-validation experiments. This was done by determining the number of true positive/false negative class predictions of sequences through a random leave-sets-out analysis (see Methods for more details).

After repeating the leave-sets-out analysis 100 times using both the traditional *k*-spectrum kernel as well as the folded *k*-spectrum kernel (with feature elimination) during the initial

training, leaving out subsets of the positive as well as negative (scrambled) CRMs for testing, the resulting number of false negatives found are shown in Fig 3B and in S7 Table. One should note that the SVM with folded *k*-spectrum performs better than the traditional SVM with *k*-spectrum method in all CRMs.

The overall performance of the classifiers are evaluated through the ROCcurves, i.e., false positive rate $\left(\frac{FP}{FP+TN}\right)$ vs. true positive rate $\left(\frac{TP}{TP+FN}\right)$ and the precision-recall curves (PR), i.e., precision (or positive predictive value) $\left(\frac{TP}{TP+FP}\right)$ vs. recall (or true positive rate). Fig 4 shows the superimposed ROC/PR curves of the 100 binary classification results, with the corresponding average Area Under the Curve scores for ROC (AUC_{avg}) and for PR ($AUCPR_{avg}$). Fig 5 shows the results corresponding to the top four panels of Fig 4 when the feature elimination procedure is applied. The folded *k*-spectrum approach clearly outperforms the contiguous *k*-spectrum in terms of *AUC* and *AUCPR* scores, with the feature elimination procedure further improving the classifier's performance.

We also compared our results to those of the repDNA algorithm [35] by generating the *k*-mers (with $k = 1, \ldots, 6$) and a set of features incorporating the gapped dependency, i.e., dinucleotide-based autocovariance (DAC) features, dinucleotide-based cross covariance (DCC) features, and pseudo dinucleotide composition (PseDNC) features. We applied 10-fold cross-validation using the binary classifier SVM, and repeated this test 100 times. The results illustrate that the folded *k*-spectrum approach also outperforms the repDNA algorithm in terms of both *AUC* and *AUCPR* scores (Fig 4).

Enriched motifs correspond to known TFBSs

For the enriched motifs found after filtering in the 9 Group 1 CRMs, we set out to gain some insight into what particular TFs these motifs may be binding *in vivo*. The improved predictive power obtained by the gapped *k*-mer models in those CRMs suggests the presence of TFs with strong nucleotide interdependencies in their binding sites. To identify these candidate TFs, we analyzed the enriched gapped *k*-mers using the insect motif database, JASPAR, along with the widely used comparison tool, Tomtom, quantifying the similarity between our enriched motifs (gapped *k*-mer sequences) and known motifs for regulatory factors [59, 60].

We analyzed all 817 enriched gapped *k*-mers found in the Group 1 CRMs, and found 136 significant hits using JASPAR (p-value $< 10^{-3}$, see S8 Table). The most prominent feature of this analysis was that a large number of the enriched gapped *k*-mers correspond to TFBSs for TFs known to regulate genes involved in early AP patterning in *Drosophila*.

TFs known to regulate AP patterning genes. The hits found by all gapped *k*-mers correspond to 29 different TFs, 14 of which are known to be involved in regulating AP patterning genes [2, 62]. These include BCD, BTD, GSC, OC, KR, TTK, KNI, TLL, HKB, H, SLP, OPA, ODD, and RUN. It is not surprising to find such a large number of the significant hits corresponding to TFs known to regulate AP patterning genes since the 127 CRMs used for the analysis are known to regulate genes that are differentially expressed along the AP-axis. However, it is worth noting that through this analysis some interesting gapped dependency patterns emerged. As an example, we focus on BCD, the TF found with the second highest number of occurrences in Group 1 CRMs.

The enriched gapped *k*-mers that were found to be significant with the known BCD motif using Tomtom were aligned and used to construct the weblogo in Fig 6B. One should note that this weblogo is very similar to JASPAR's BCD motif (Fig 6A), in which the gaps significantly align to the 4th base and the (flanking) 7th base of the JASPAR motif. Although the majority (7) of these aligned enriched gapped *k*-mers have the contiguous 5-mer model

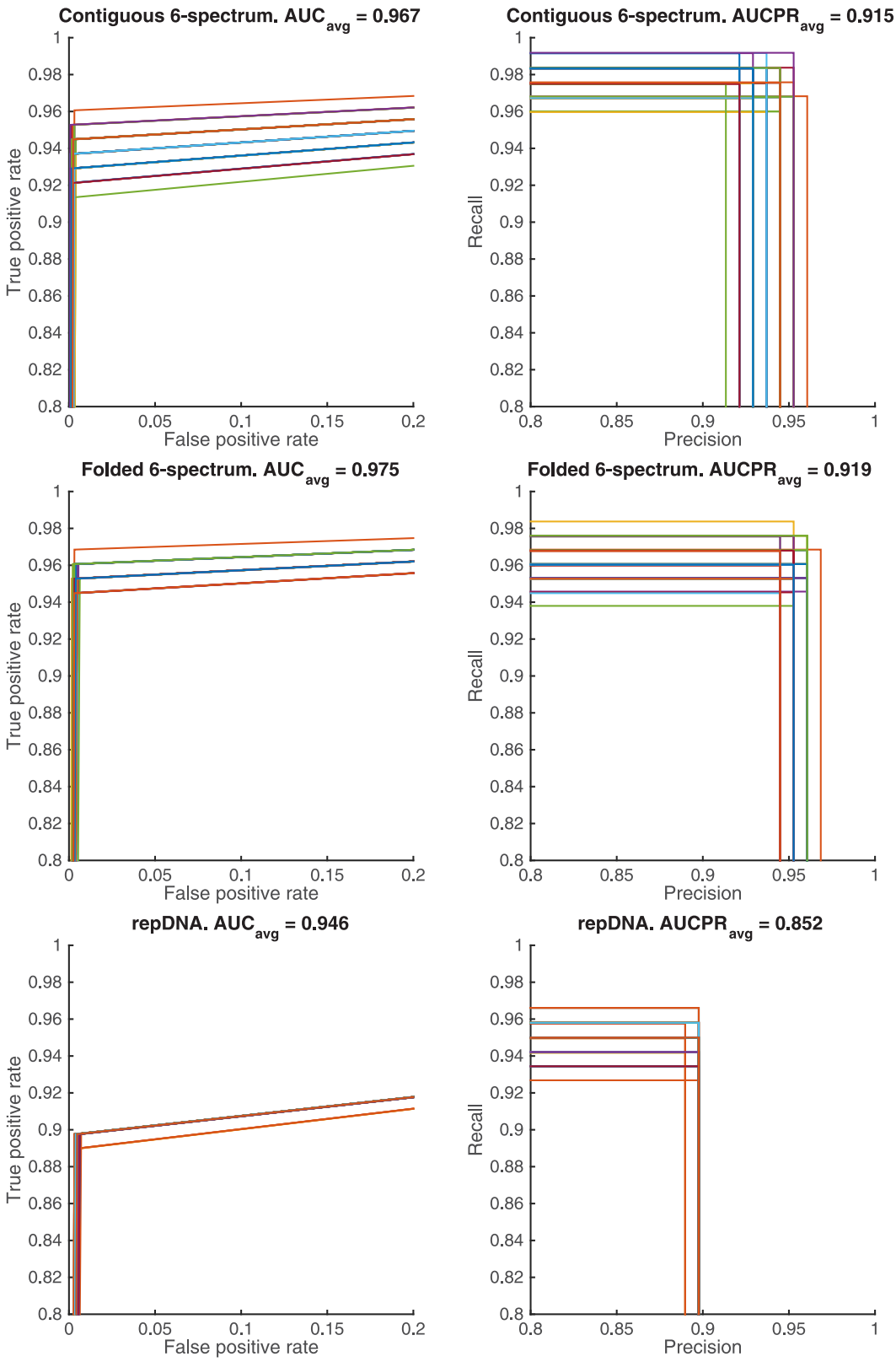

Fig 4. Superimposed performance curves, ROC (left) and PR (right), of the 100 cross-validation tests by SVM learning with contiguous *k*-spectrum kernel (top), folded *k*-spectrum kernel (middle), and repDNA features (lower). Note that many of the ROC curves overlap due to the very small number of false positives found. Different colors represent different cross-validation results.

https://doi.org/10.1371/journal.pone.0185570.g004

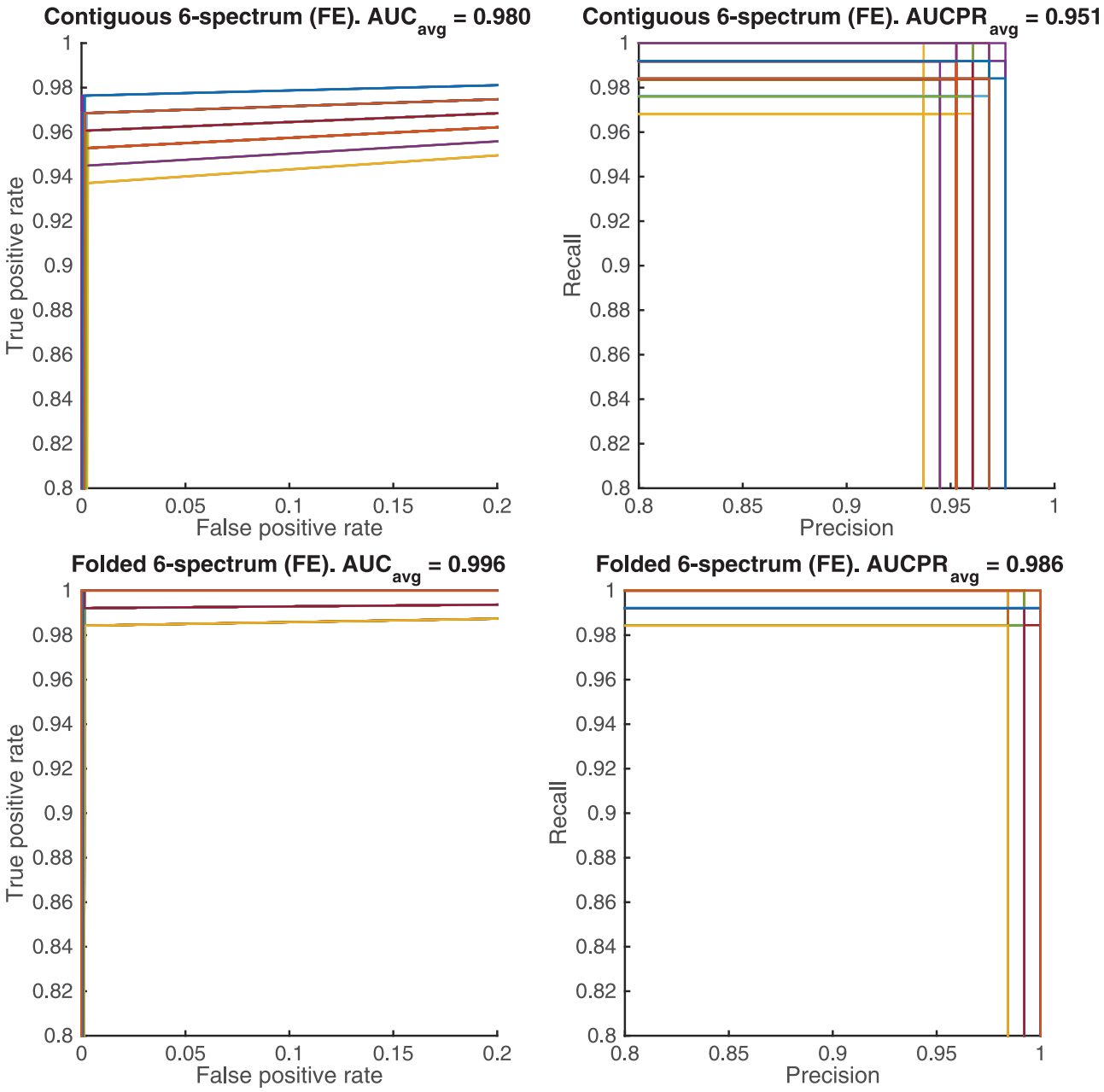

Fig 5. Superimposed performance curves, ROC (left) and PR (right), of the 100 cross-validation tests by using the feature elimination (FE) in SVM learning with contiguous *k*-spectrum kernel (above) vs. folded *k*-spectrum kernel (below). Note that many of the ROC curves overlap due to the very small number of false positives found. Different colors represent different cross-validation results.

https://doi.org/10.1371/journal.pone.0185570.g005

(*Nαααααα*), the remainder of them (3) have a consistent appearance of the gap in the 4th base, suggesting a non-adjacent (gapped) dependency between nucleotides 2 and 3, and 5 and 6.

One possible explanation for this gapped dependency may be the three-dimensional shape of the DNA when bound by BCD. BCD is a homeodomain TF, and thus is thought to mediate DNA binding through contact primarily with nucleotides in the major groove using the homeodomain recognition helix, although there is evidence of additional DNA contact in the minor groove via the relatively unstructured N-terminal domain [63–65]. Therefore, nucleotides in

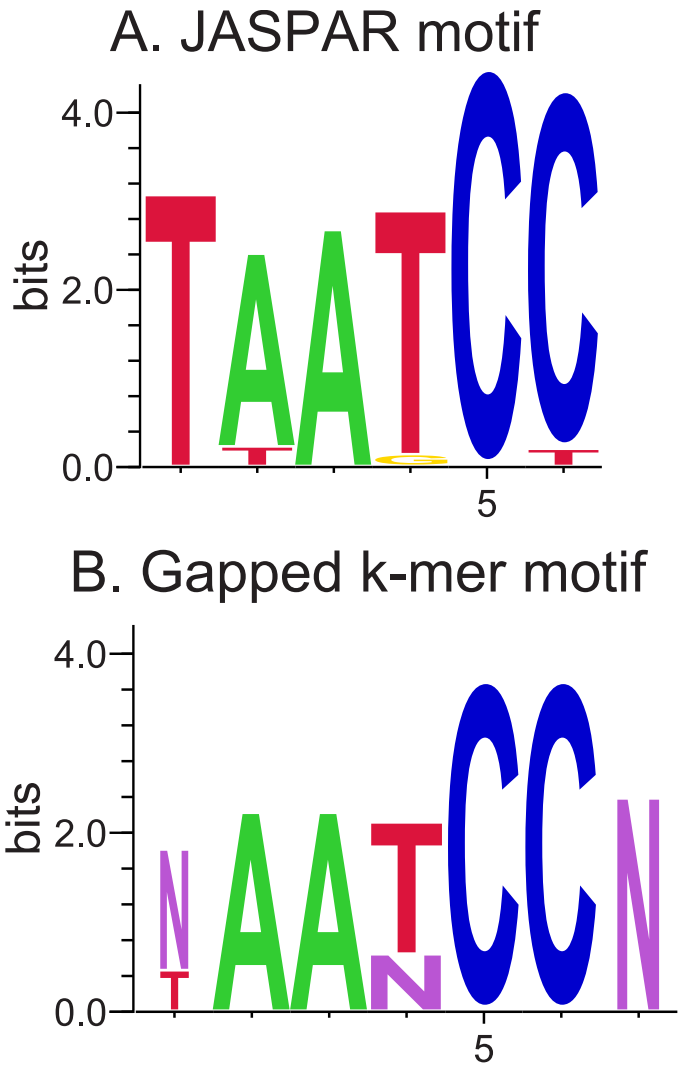

Fig 6. Weblogos of the JASPAR motif (A) and the enriched gapped *k*-mer motifs (B) corresponding to the BCD TFBSs. Note that in the weblogos, A, C, G, and T represent the corresponding nucleotides, while purple N's in (B) represent gaps introduced by the enriched gapped *k*-mers.

https://doi.org/10.1371/journal.pone.0185570.g006

the transition between the major and minor groove may not be as important when binding. When the BCD binding sites in the JASPAR database, originally identified using bacterial-1 hybrid experiments [62], were analyzed using the newly developed and validated high-throughput approach, DNAshape, [66] they were predicted to contain a significantly smaller minor groove width estimated on the 4th base [67]. We believe the gap we observe in the 4th base (and the entailing dependency between the surrounding nucleotides) may be due to this reduced minor groove width, signifying the nucleotide position where the transition occurs from the major to minor groove.

Discussion

Approaches to binding site prediction from DNA sequence data have been developed, and improved upon for decades. The development and implementation of the novel approach laid out in this manuscript has led us to some very interesting conclusions. Most striking, and

probably most important to the general bioinformatic community, is the illustrated advantages that this new approach, which incorporates gapped dependency, has shown over existing SVM approaches. To highlight the success of this approach, note that 126 of the 127 CRMs tested were correctly identified as belonging to the 'positive sequence set' by this new algorithm, while 9 of those were not correctly identified by the previous contiguous *k*-mer approach, and the only remaining CRM was incorrectly identified by both approaches. Thus, our new algorithm outperformed the previous algorithm over the complete set of 127 CRMs (Fig 3B). We also found, through a rigorous set of cross-validation tests, that our novel approach outperformed the previous approach in terms of the *AUC* measurements (average of 0.996 vs. 0.98, Fig 5), as well as in terms of the *AUCPR* measurements (average of 0.986 vs. 0.951, Fig 5).

Beyond showing the advantage this approach holds over previous SVM approaches, we have also illustrated its utility on the specific set of CRMs tested, validating the method and raising new biological hypotheses regarding the nature of DNA-protein binding. The 127 CRMs chosen for this study belong to 114 genes which exhibit differential expression patterns along the AP-axis in early *Drosophila* embryos. When comparing the gapped *k*-mers found to be enriched using our algorithm to those in the insect motif database, JASPAR, we found that 14 of the 29 TFs found were indeed known to be involved in regulating AP patterning genes. Although not all TFs found were known to be involved in the regulation of AP patterning genes, the appearance of many of these other TFs can still be explained in a reasonable biological context. For example, BRK and MAD are thought to compete for binding sites effecting Dpp signaling events in early *Drosophila* development, and our results support this as we have found them enriched on the same enhancers. We have also found enrichment of CTCF binding site sequences within these 127 known CRMs, which has been shown to be involved in enhancer-promoter looping. These results all lead us to believe that the motifs we are identifying, in many cases, likely represent true TF binding sites.

Conclusion

This study has introduced a novel approach to motif identification, which builds upon previous approaches in machine learning to allow for a less biased approach to binding site discovery. We have shown its ability to predict TF binding sites on a set of 127 CRMs, and have offered some insight into the molecular basis for its success. In the future, it will be very interesting to see similar analyses performed on various other sets of sequences, possibly including sequences from different species and different time points in development. Such studies could help in answering a deeper biological question of whether universal rules exist governing binding site preference, strength, and flexibility.

Supporting information

S1 Table. Classification performance of the (contiguous) k-spectrum kernel vs. folded k-spectrum kernel (Fig 3).
(XLS)

S2 Table. Gapped k-mers enriched in the Group 1 CRMs.
(XLS)

S3 Table. Gapped k-mers enriched in each of the 127 CRMs.
(XLS)

S4 Table. The gapped k-mer models enriched in the Group 1 CRMs with total counts and cumulative weights of the corresponding motifs from S2 Table.
(XLS)

S5 Table. The gapped k-mers enriched in the Group 1 CRMs by using the feature elimination procedure in the SVM methods.
(XLS)

S6 Table. The gapped k-mer models enriched in the Group 1 CRMs by using the feature elimination procedure in the SVM methods.
(XLS)

S7 Table. Classification performance of the (contiguous) k-spectrum kernel vs. folded k-spectrum kernel with feature elimination (FE) procedure (Fig 4).
(XLS)

S8 Table. Motif comparison results with the enriched gapped k-mers in the Group 1 CRMs (from S5 Table).
(XLS)

S1 File. Folded k-spectrum program. Matlab implementations of the algorithms described in the Methods and Results sections.
(ZIP)

Acknowledgments

We thank Ahmet Ay for kindly introducing the authors of this study.

Author Contributions

Conceptualization: Abdulkadir Elmas, Xiaodong Wang, Jacqueline M. Dresch.

Data curation: Abdulkadir Elmas.

Formal analysis: Abdulkadir Elmas.

Funding acquisition: Jacqueline M. Dresch.

Investigation: Xiaodong Wang, Jacqueline M. Dresch.

Methodology: Abdulkadir Elmas, Jacqueline M. Dresch.

Project administration: Xiaodong Wang, Jacqueline M. Dresch.

Resources: Abdulkadir Elmas.

Software: Abdulkadir Elmas.

Supervision: Xiaodong Wang, Jacqueline M. Dresch.

Validation: Abdulkadir Elmas, Xiaodong Wang, Jacqueline M. Dresch.

Visualization: Abdulkadir Elmas.

Writing – original draft: Abdulkadir Elmas, Xiaodong Wang, Jacqueline M. Dresch.

Writing – review & editing: Abdulkadir Elmas, Xiaodong Wang, Jacqueline M. Dresch.

References

1. Borok M, Tran D, Ho M, RA D. Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. Development. 2010; 137(1):5–13. https://doi.org/10.1242/dev.036160 PMID: 20023155

2. Starr M, Ho M, Gunther E, Tu Y, Shur A, Goetz S, et al. Molecular dissection of cis-regulatory modules at the Drosophila bithorax complex reveals critical transcription factor signature motifs. Dev Biol. 2011; 359(2):290–302. https://doi.org/10.1016/j.ydbio.2011.07.028 PMID: 21821017
3. Crocker J, Tamori Y, Erives A. Evolution Acts on Enhancer Organization to Fine-Tune Gradient Threshold Readouts. PLoS Biol. 2008; 6(11):2576–2587. https://doi.org/10.1371/journal.pbio.0060263
4. Martinez CA, Barr K, Kim AR, Reinitz J. A synthetic biology approach to the development of transcriptional regulatory models and custom enhancer design. Methods. 2013; 62(11):91–98. https://doi.org/10.1016/j.ymeth.2013.05.014 PMID: 23732772
5. Waterman MS, Joyce J, Eggert M. Computer alignment of sequences. Phylogenetic Analysis of DNA Sequences. 1991; p. 59–72.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. A basic local alignment search tool. Journal of Molecular Biology. 1990; 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712
7. Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1995. Nucleic Acids Res. 1996; 24(1):189–96. https://doi.org/10.1093/nar/24.1.189 PMID: 8594577
8. Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN. The PRINTS protein fingerprint database in its fifth year. Nucleic Acids Res. 1998; 26(1):304–308. https://doi.org/10.1093/nar/26.1.304 PMID: 9399860
9. Krogh A, Brown M, Mian I, Sjolander K, Haussler D. Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biol. 1994; 235:1501–1531. https://doi.org/10.1006/jmbi.1994.1104
10. Eddy SR. Multiple alignment using hidden Markov models. Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology. 1995; p. 114–120.
11. Zellers RG, Drewell RA, Dresch JM. MARZ: an algorithm to combinatorially analyze gapped *n*-mer models of transcription factor binding. BMC Bioinf. 2015; 16:30. https://doi.org/10.1186/s12859-014-0446-3
12. Dresch JM, Zellers RG, Bork DK, Drewell RA. Nucleotide interdependency in transcription factor binding sites in the Drosophila genome. Gene Regulation and Systems Biology. 2016;in press. https://doi.org/10.4137/GRSB.S38462 PMID: 27330274
13. Stormo G, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Res. 1982; 10(9):2997–3011. https://doi.org/10.1093/nar/10.9.2997 PMID: 7048259
14. Staden R. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 1984; 12(1Part2):505–519. https://doi.org/10.1093/nar/12.1Part2.505 PMID: 6364039
15. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. Journal of Molecular Biology. 1987; 193(4): 723–750. https://doi.org/10.1016/0022-2836(87)90354-8 PMID: 3612791
16. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics. 1998; 14(1):48–54. https://doi.org/10.1093/bioinformatics/14.1.48 PMID: 9520501
17. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999; 15(7-8):563–577. https://doi.org/10.1093/bioinformatics/15.7.563 PMID: 10487864
18. Man TK, Stormo GD. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res. 2001; 29: 2471–2478. https://doi.org/10.1093/nar/29.12.2471 PMID: 11410653
19. Benos PV, Lapedes AS, Stormo GD. Probabilistic code for DNA recognition by proteins of the EGR family. Journal of Molecular Biology. 2002; 323:701–727. https://doi.org/10.1016/S0022-2836(02)00917-8 PMID: 12419259
20. Lassig M. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics. 2007; 8(Suppl 6):S7. https://doi.org/10.1186/1471-2105-8-S6-S7 PMID: 17903288
21. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009; 324(5935):1720–1723. https://doi.org/10.1126/science.1162327 PMID: 19443739
22. Siddharthan R. Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. PLoS ONE. 2010; 5(3):e9722. https://doi.org/10.1371/journal.pone.0009722 PMID: 20339533
23. Annala M, Laurila K, Lahdesmaki H, Nykter M. A linear model for transcription factor binding affinity prediction in protein binding microarrays. PLoS One. 2011; 6(5):e20059. https://doi.org/10.1371/journal.pone.0020059 PMID: 21637853

24. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press; 2000.
25. Vapnik V. Statistical Learning Theory. Springer; 1998.
26. Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. Journal of Computational Biology. 2000; 7(1/2):95–114. https://doi.org/10.1089/10665270050081405 PMID: 10890390
27. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247(4):536–540. https://doi.org/10.1016/S0022-2836(05)80134-2 PMID: 7723011
28. Liao L, Noble WS. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. Proceedings of the sixth annual international conference on Computational biology. 2002; p. 225–232.
29. Leslie C, Eskin E, Noble WS. The spectrum kernel: A string kernel for SVM protein classification. Pac Symp Biocomput. 2002; 7:564–575.
30. Leslie C, Eskin E, Weston J, Noble WS. Mismatch string kernels for SVM protein classification. Advances in Neural Information Processing Systems. 2003; 15:1417–1424.
31. Leslie C, Eskin E, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics. 2004; 4:467–476. https://doi.org/10.1093/bioinformatics/btg431
32. Mathelier A, Wasserman WW. The Next Generation of Transcription Factor Binding Site Prediction. J Bioinform Comput Biol. 2013; 9(9):e1003214.
33. Magbanua JP, Runneburger E, Russell S, White R. A general pairwise interaction model provides an accurate description of in vivo transcription factor binding sites. PLoS One. 2014; 9(6):e99015. https://doi.org/10.1371/journal.pone.0099015
34. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLOS Computational Biology. 2014; 10(7):1–15. https://doi.org/10.1371/journal.pcbi.1003711
35. Liu B, Liu F, Fang L, Wang X, Chou KC. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics. 2015; 31(8):1307–1309. https://doi.org/10.1093/bioinformatics/btu820 PMID: 25504848
36. Lee D, Karchin R, Beer M. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011; 21(12):2167–80. https://doi.org/10.1101/gr.121905.111 PMID: 21875935
37. Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, et al. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. PLoS Comput Biol. 2014; 6:e1003677. https://doi.org/10.1371/journal.pcbi.1003677
38. Liu B, Fang L, Wang S, Wang X, Li H, Chou KC. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J Theor Biol. 2015; 385:153–159. https://doi.org/10.1016/j.jtbi.2015.08.025 PMID: 26362104
39. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015; 43(W1):W65–71. https://doi.org/10.1093/nar/gkv458 PMID: 25958395
40. Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, et al. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One. 2014; 9(9):e106691. https://doi.org/10.1371/journal.pone.0106691 PMID: 25184541
41. Zhang H, Zhu L, Huang DS. WSMD: weakly-supervised motif discovery in transcription factor ChIP-seq data. Sci Rep. 2017; 7(1):3217. https://doi.org/10.1038/s41598-017-03554-7 PMID: 28607381
42. Zhu L, Zhang H, Huang DS. LMMO: A Large Margin Approach for Refining Regulatory Motifs. IEEE/ACM Trans Comput Biol Bioinform. 2017; https://doi.org/10.1109/TCBB.2017.2691325
43. Stringham JL, Brown AS, Drewell RA, Dresch JM. Flanking sequence context-dependent transcription factor binding in early *Drosophila* development. BMC Bioinf. 2013; 14:298. https://doi.org/10.1186/1471-2105-14-298
44. Gallo S, Gerrard D, Miner D, Simich M, Des Soye B, Bergman C, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Res. 2011; 39:D118–D123. https://doi.org/10.1093/nar/gkq999 PMID: 20965965
45. Batista GEAPA, Prati RC, Monard MC. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explor Newsl. 2004; 6(1):20–29. https://doi.org/10.1145/1007730.1007735

46. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Int Res. 2002; 16(1):321–357.
47. Wu X, Srihari R. Incorporating Prior Knowledge with Weighted Margin Support Vector Machines. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'04. New York, NY, USA: ACM; 2004. p. 326–333. Available from: http://doi.acm.org/10.1145/1014052.1014089.
48. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A. 2000; 97(1):262–267. https://doi.org/10.1073/pnas.97.1.262 PMID: 10618406
49. Joachims T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Norwell, MA, USA: Kluwer Academic Publishers; 2002.
50. Veropoulos K, Campbell C, Cristianini N. Controlling the Sensitivity of Support Vector Machines. In: Proceedings of the International Joint Conference on AI; 1999. p. 55–60.
51. Zadrozny B, Langford J, Abe N. Cost-sensitive learning by cost-proportionate example weighting. In: Third IEEE International Conference on Data Mining; 2003. p. 435–442.
52. Wang L, Gao Y, Chan KL, Xue P, Yau WY. Retrieval with knowledge-driven kernel design: an approach to improving SVM-based CBIR with relevance feedback. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. vol. 2; 2005. p. 1355–1362 Vol. 2.
53. Mathew J, Luo M, Pang CK, Chan HL. Kernel-based SMOTE for SVM classification of imbalanced datasets. In: IECON 2015—41st Annual Conference of the IEEE Industrial Electronics Society; 2015. p. 001127–001132.
54. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994; 2:28–36. PMID: 7584402
55. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004; 5:276–287. https://doi.org/10.1038/nrg1315 PMID: 15131651
56. Turatsinze J, Thomas-Chollier M, Defrance M, van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc. 2008; 3(10):1578–88. https://doi.org/10.1038/nprot.2008.97 PMID: 18802439
57. Hardison R, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet. 2012; 13(7):469–83. https://doi.org/10.1038/nrg3242 PMID: 22705667
58. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol. 2013; 31(2):126–134. https://doi.org/10.1038/nbt.2486 PMID: 23354101
59. Mathelier A, Zhao X, Zhang A, Parcy F, Worsley-Hunt R, Arenillas D, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2013; 42:D142–7. https://doi.org/10.1093/nar/gkt997 PMID: 24194598
60. Gupta S, Stamatoyannopoulos J, Bailey T, Noble W. Quantifying similarity between motifs. Genome Biol. 2007; 8(2):R24. https://doi.org/10.1186/gb-2007-8-2-r24 PMID: 17324271
61. McQuilton P, Pierre SES, Thurmond J, the FlyBase Consortium. FlyBase 101? the basics of navigating FlyBase. Nucleic Acids Res. 2012; 40:D706–D714. https://doi.org/10.1093/nar/gkr1030 PMID: 22127867
62. Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res. 2008; 36(8):2547–2560. https://doi.org/10.1093/nar/gkn048 PMID: 18332042
63. Desplan C, Theis J, O'Farrell P. The sequence specificity of homeodomain-DNA interaction. Cell. 1988; 54(7):1081–1090. https://doi.org/10.1016/0092-8674(88)90123-7 PMID: 3046753
64. Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, et al. Homeodomain-DNA recognition. Cell. 1994; 78(2):211–223. https://doi.org/10.1016/0092-8674(94)90292-5 PMID: 8044836
65. Baird-Titus J, Clark-Baldwin K, Dave V, Caperelli C, Ma J, Rance M. The solution structure of the native K50 Bicoid homeodomain bound to the consensus TAATCC DNA-binding site. J Mol Biol. 2006; 356(5):1137–1151. https://doi.org/10.1016/j.jmb.2005.12.007 PMID: 16406070
66. Zhou T, Yang L, Lu Y, Dror I, Dantas Machado A, Ghane T, et al. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. 2013; 41: W56–62. https://doi.org/10.1093/nar/gkt437 PMID: 23703209
67. Yang L, Zhou T, Dror I, Mathelier A, Wasserman WW, Gordan R, et al. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 2014; 42(Database issue):D148–55. https://doi.org/10.1093/nar/gkt1087 PMID: 24214955