Selecting Predictive Features for Recognition of Hypersensitive Sites of Regulatory Genomic Sequences with an Evolutionary Algorithm

Uday Kamath
ukamath@gmu.edu

Kenneth A De Jong
kdejong@gmu.edu

Amarda Shehu*
amarda@gmu.edu

Department of Computer Science
George Mason University
Fairfax, VA 22030

ABSTRACT

This paper proposes a method to improve the recognition of regulatory genomic sequences. Annotating sequences that regulate gene transcription is an emerging challenge in genomics research. Identifying regulatory sequences promises to reveal underlying reasons for phenotypic differences among cells and for diseases associated with pathologies in protein expression. Computational approaches have been limited by the scarcity of experimentally-known features specific to regulatory sequences. High-throughput experimental technology is finally revealing a wealth of hypersensitive (HS) sequences that are reliable markers of regulatory sequences and currently the focus of classification methods. The contribution of this paper is a novel method that combines evolutionary computation and SVM classification to improve the recognition of HS sequences. Based on experimental evidence that HS regions employ sequence features to interact with enzymes, the method seeks motifs to discriminate between HS and non-HS sequences. An evolutionary algorithm (EA) searches the space of sequences of different lengths to obtain such motifs. Experiments reveal that these motifs improve recognition of HS sequences by more than 10% compared to state-of-the-art classification methods. Analysis of these motifs reveals interesting insight into features employed by regulatory sequences to interact with DNA-binding enzymes.

Track: Bioinformatics, Computational, Systems and Synthetic Biology

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life And Medical Sciences—*Biology and genetics*; I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*[Knowledge acquisition]

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GECCO'10, July 7–11, 2010, Portland, Oregon, USA.
Copyright 2010 ACM 978-1-4503-0072-8/10/07 ...$10.00.

General Terms

Algorithms

Keywords

DNase I Hypersensitive Sites, Evolutionary Algorithms, Support Vector Machines

1. INTRODUCTION

The human genome project resulted in the annotation of $\sim 25,000$ protein-encoding genes [7]. Cells of different tissue were found to have the same set of genes yet display significant phenotypical differences. Experimental evidence now amounts that cell development, proliferation, apoptosis, aging, differentiation, and even certain pathologies, depend on the spatial and temporal expression of genes [28, 17].

Protein expression in cells is controlled by regulating which genes are transcribed into mRNA. DNA-binding proteins known as transcription factors bind specific DNA sequences known as regulatory elements to activate or repress transcription of following genes. The DNA footprint of this set of factors is called a cis-regulatory module and consists of a set of factor-binding sites located a few hundred bases upstream of the gene being regulated [8, 19].

Understanding what governs expression patterns and their alterations in eukaryotic cells relies on identifying regulatory elements associated with an annotated gene. Full annotation of the human genome with regulatory elements is an emerging challenge in genomics research [28, 17]. Evidence has emerged over the past twenty years that the eukaryotic cell hides regulatory elements by wrapping and packaging DNA into chromatin fibers [41, 14, 12, 13].

DNA-modifying enzymes like non-specific endonuclease DNase I disrupt chromatin architecture to expose regulatory sequences. Experimental research shows that DNA sites that are hypersensitive to DNase I, referred to as HS sites, are reliable markers of regulatory elements. HS sites are found to be co-localized with regulatory elements in all known regulatory classes (transcriptional enhancers, promoters, silencers, insulators, splicing regulators, and chromatin boundary elements) [41, 14, 25, 4, 34, 28, 17].

Until the discovery of HS sites, wet-lab identification of regulatory elements relied on detecting transcription factor-binding sites [2]. Mapping HS sites now underpins the discovery of most regulatory elements in the human genome [34].

Bolstered by experimental success, mapping HS sites is considered the gold approach for discovering regulatory elements. Identification of these sites promises to reveal almost all classical regulatory sequences and dramatically accelerate the functional annotation of the entire human genome.

Efforts to detect HS sites and employ them as generic markers of regulatory sequences span experimental and computational approaches. Early experimental methods to visualize HS sites relied on indirect approaches and laborious localization of activating sequences [41, 25, 37]. Recent years have seen a burst of high-throughput experimental methods that are recovering a large number of HS sequences [34, 10]. Comprehensive annotation, however, is expected to require substantial efforts. A combination of computational and experimental approaches may be needed for a rapid annotation of the entire human genome [30].

The large number of discovered HS sequences have created the opportunity to develop computational methods that can learn to recognize these sequences and assist in genome-wide annotation of regulatory elements. Early approaches, which suffered from the scarcity of available HS sequences, focused on identification and combinatorial analysis of short sequence motifs presumed to represent transcription factor-binding sites in regulatory sequences. The motifs were often obtained from experiment, analysis of upstream regions of co-regulated genes [35, 1], or employment of phylogenetic data [32]. Performance was generally poor [38] and did not improve until a large repository of known HS sequences was made available to machine learning approaches.

Recent computational approaches to recognition of HS sequences employ Support Vector Machines (SVM) [30]. In computational biology, SVMs have been applied to prediction of protein localization sites [15], DNA translation start sites [43], DNA splice sites [42, 21, 20], and many more problems (cf. to [29]). In particular, work in [30] shows that a sequence-based SVM classifier learns to recognize HS sequences with accuracy more than 80%. The SVM in [30] enumerates all possible subsequences of fixed length and uses their occurrences in input data as features through which to discriminate between HS and non-HS sequences.

HS sequences are believed to contain complex features that facilitate recognition by specific sets of DNA-binding proteins interacting cooperatively over short distances of 150-250 base pairs [12, 36]. While it is not known *a priori* whether recognizing such features is computationally tractable, a reasonable hypothesis is that short motif-like sequence features can capture differences between HS and non-HS sequences. In the absence of *a priori* knowledge of such motifs, the SVM in [30] employs a brute-force approach by considering all finite subsequences of fixed length as potential motifs. However, even if features in known HS sequences are captured by some subsequences, the noise from the rest of the subsequences may adversely impact the performance of a classifier. Moreover, a brute-force approach sheds little insight into which subsequences may be employed by HS sequences to interact with DNA-binding proteins.

This paper pursues the hypothesis that motifs, short subsequences of finite length that capture and maximize differences between HS and non-HS sequences, exist and can be employed to improve the accuracy of supervised learning methods. Our goal is to identify such motifs with an evolutionary computation framework. We develop an Evolutionary Algorithm (EA) that searches the space of possible motifs for those that maximize differences between known HS and non-HS sequences. Comprehensive testing of the method reveals that better motifs are obtained when employing an "island model" to independently (and in parallel) evolve motifs of different lengths.

A systematic analysis measures improvements in classification accuracy when employing the obtained motifs to either replace or enhance subsequences of fixed length used as features in an SVM-based classifier. Each experimental setting trains an SVM classifier on 280 HS sequences annotated throughout the human genome and 737 confirmed non-HS sequences of equivalent length. The analysis shows that the motifs obtained by our EA algorithm afford better classification accuracy, by as much as 10% over a setting where not motifs but all subsequences of fixed length are used as features in the SVM. The results make the case that short motifs can be designed to better capture the sequence features recognized by DNA-binding proteins in HS sequences and improve the performance of machine learning techniques. From a methodology point of view, our work proposes the combination of evolutionary-computing approaches with statistical machine learning approaches to improve recognition of HS sequences.

The rest of this paper is organized as follows. A brief summary of related work is provided in section 1.1. Our method is described in section 2. Results follow in section 3. The article concludes with a discussion in section 4.

1.1 Related Work

In the following we briefly summarize SVMs and related efforts on SVM-based recognition of HS sequences.

1.1.1 Support Vector Machines

SVMs have a solid theoretical foundation in statistical learning theory [39]. They have wide applicability in binary classification, in part because of their successful application to non-vector data such as text, graphs, and strings. They do so by requiring the user to define a mapping into a feature space. Then, given n labeled training data in feature vector form, an SVM classifier first transforms the data into an internal vector space, typically of higher dimensionality. The transformation is chosen so that the positive and negative examples are linearly separable by a hyperplane in the higher-dimensional space. Then, predicting the label of an unlabeled data point involves determining on which side of the hyperplane the point lies.

The success of this approach depends on both the choice of the feature space used and the internal transformation, the "kernel function," used. Well-known kernel functions include the Polynomial, Radial Basis, Gaussian Radial, and Sigmoid functions [3]. The particular choice of a kernel is problem-specific and generally determined experimentally. As discussed in the next section, the experiments reported here use a Radial Basis kernel function (RBF).

The focus of this research is on improving the feature space used as input to an SVM. In our case, HS sites (positive examples) and non-HS sites are variable-length sequences of nucleotides, and thus easily described as strings drawn from the familiar {A,C,G,T} alphabet.

The most recent successful SVM application to recognition of HS sequences employs a "spectrum" feature space. Spectrum features are all finite-length sequences that can be generated from an alphabet Σ [24, 5]. For a fixed k, a

k-spectrum is the set of $d = |\Sigma|^k$ features that correspond to all strings of length k (k-mers). A data string S of length L is mapped into this d-dimensional feature space by simply recording the normalized occurrence of each k-mer in S. The normalization is with respect to the number of times any k-mer appears in the sequence. For instance, the sequence ACGT contains three 2-mers AC, CG, and GT. The $k = 2$ spectrum representation of this sequence is the 16-element vector with 0.33 entries corresponding to the 2-mers above and 0 entries corresponding to all others.

The sense one gets is that this is a rather crude method for selecting discriminating features for an SVM. The motivation for this research is the belief that SMV classification could be enhanced by a more sophisticated feature space based on motifs evolved by an appropriately designed evolutionary algorithm. These motifs may additionally shed light into what sequence-based features are employed by HS sequences to initiate recognition and interaction with DNA-binding proteins.

2. METHODS

We propose an evolutionary algorithm (EA) to design motifs, short subsequences of finite length, that capture and maximize differences between HS and non-HS sequences for classification purposes. The validation of the usefulness of these motifs as opposed to brute-force feature spaces defined by all subsequences of finite length is performed in the context of an SVM-based classifier.

Unlike work in [30] which employs linear kernel functions, our SVM classifier employs an RBF. Recent work suggests that RBFs perform better than polynomial kernel functions [15]. While significant work is devoted to determining optimal kernel functions for specific problem domains [31, 27, 26], the focus in this work is on determining optimal motifs that capture important features of HS sequences.

The EA we designed searches the space of finite-length sequences to find those that best discriminate between HS and non-HS sequences. The sequences employed here have variable length $l \in \{6, \ldots, 12\}$. The reason for searching over sequences of variable length is that we do not a priori know the length needed to capture features intrinsic to HS sequences. The choice of the lower bound reflects findings in [30], where 6-mers as spectrum features confer high accuracy to an SVM classifier of HS and non-HS sequences.

We point out that, while the motifs are intended to be used in an SVM classifier, the objective fitness function in our EA does not employ the SVM classifier. Since our method evolves many motifs for thousands of generations, employing an SVM to measure the fitness of each generated motif is not practical. For this reason, we employ a simple fitness function that depends on occurrences of a particular motif in HS and non-HS sequences (detailed below).

Our detailed description of the method focuses on the EA we use to design better motifs (i.e., better features) by which to discriminate between HS and non-HS sequences in the context of powerful classifiers like SVMs. We describe the EA search for such motifs after a brief summary of the data set and the performance measure and cross-validation employed to quantify the accuracy in the SVM-based recognition of HS sequences.

2.1 Data

The input data set employed in the fitness function of our EA and the training and cross-validation of the SVM consists of 280 experimentally-established erythroid HS sequences and 737 non-HS sequences. This data set was obtained from noble.gs.washington.edu/proj/hs. The HS sequences were identified from throughout the human genome through a novel experimental methodology that identifies HS sites employing cloning and in-vivo activity of K562 erythroid cells [34]. The non-HS sequences, collected (and distributed proportionally) from throughout the human genome, were not hypersensitive when tested in the same cell type. We point out that both 280 HS and 737 non-HS sequences have similar average lengths of 242 nucleotides.

2.2 Performance measure

We employ receiver operating characteristic (ROC) curves to measure and compare the quality of an SVM classifier [16]. A trained SVM outputs a ranked list of HS sequences, ordered from the most to least confident. If one varies a threshold from the top to the bottom of the list, the rate of true and false positives change with the threshold. An ROC curve plots the true positive rate as a function of the false positive rate as the threshold changes. The ROC score refers to the area under the curve. While random ranking is expected to yield a score of ~ 0.5, the ROC score reaches 1 if the SVM correctly places all of the HS sequences above the threshold.

2.3 Cross-validation

The performance of the SVM is tested via 10-fold cross-validation on the total set of 1017 sequences (280 HS and 737 non-HS). The set is randomly divided into 10 subsets of equal size. The SVM is trained on 90% of the subsets and tested on the 10% held out. The area under the ROC curve is reported as an average over the 10-fold validations. The obtained mean scores are employed in our comprehensive analysis to compare an SVM with spectrum feature frequencies and an RBF kernel, spectrum feature presence or absence and RBF, our EA-obtained motifs frequencies and RBF, EA-obtained motifs presence or absence and RBF, a combination of EA-obtained motifs and spectrum feature frequencies and RBF, and a combination of EA-obtained motifs and spectrum feature presence or absence and RBF. Average ROC scores range from $77 - 88\%$ (see section 3).

2.4 Evolutionary Algorithm

We developed a method that uses an EA to select motifs that best discriminate between HS and non-HS sequences. Our choice of an EA-based approach is a reflection of the size and complexity of the motif space being explored [33, 9, 22]. The proposed EA-based method yields DNA motifs to either replace or assist spectrum features in discriminating HS from non-HS sequences. In the work reported here, the motif alphabet is the IUPAC code [6], and motifs are defined as sequences of variable length $l \in \{6, \ldots, 12\}$ generated from this alphabet.

Our EA searches this space of candidate motifs using a $(\mu + \lambda)$-ES-style EA. In the experiments reported here, we used $\mu = 100$ parents and $\lambda = 25$ offspring, beginning with 100 randomly generated motifs. In each generation, 25 offspring are produced from (uniform) randomly selected parents using both mutation and crossover operators. Truncation selection is employed to determine which 100 of the 125 will survive as the next generation of parents.

In the experiments reported in section 3, an upper bound

5000 generations was used, although convergence in the top fitness scores was generally obtained within 500 generations. The top 25 motifs of the final population are finally evaluated for their effectiveness as features in SVM classification.

2.4.1 Representation

The individuals in our EA are variable-length strings generated from the IUPAC code for DNA sequences [6]. Our choice to vary the length from 6-mers to 12-mers reflects the lack of *a priori* information on the length of optimal motifs and the findings in [30] that no shorter than 6-mers are needed to achieve more than 80% classification accuracy with an SVM. We point out that the spectrum features employed in the SVM classification in [30] are 6-mers on the 4-letter alphabet of DNA.

Symbol	Meaning	Description Origin
G	G	**G**uanine
A	A	**A**denine
T	T	**Thymine**
C	C	**C**ytosine
R	G or A	pu**R**ine
Y	T or C	p**Y**rimidine
M	A or C	a**M**ino
K	G or T	**K**etone
S	G or C	**S**trong interaction
W	A or T	**W**eak interaction
H	A or C or T	**H** follows G in alphabet
B	G or T or C	**B** follows A in alphabet
V	G or C or A	**V** follows U in alphabet
D	G or A or T	**D** follows C in alphabet
N	G or A or T or C	a**N**y

Table 1: IUPAC code is adapted from [6].

The alphabet in our case was the larger IUPAC code detailed in Table 1, which contains characters that represent more than one nucleotide. These symbols employ ambiguous pattern-like designations, which allow motifs in which specific positions are not constrained to specific nucleotides but rather to a group of nucleotides with shared properties. In this way, the motifs offered by the EA after its final iteration may reveal interesting insight into what sequence features are employed by HS sequences to recognize and interact with DNA-binding proteins.

2.4.2 Offspring Generation

Given a current population of individuals, the EA creates an offspring population through two basic operators, mutation and crossover. The EA used in this research gives equal probability to each of the operators. That is, given a current population of individuals, there is equal chance of applying either mutation or crossover to generate a new offspring.

If the mutation operator is chosen to create a new offspring, any of the individuals in the current parent population have equal probability of being selected. Once an individual is chosen for mutation, any of its symbols has equal probability of being mutated into any of the other symbols of the IUPAC code shown in Table 1.

The crossover operator combines the genetic code of two parents to produce an offspring. Any pair of individuals has equal probability of selection for crossover. Once two individuals are chosen as parents, their genetic material is combined to produce an offspring. While our implementation allows two-point or uniform crossover [9], one-point crossover was employed to obtain the results in section 3.

2.4.3 Fitness Function

The fitness function evaluates individuals in a population in order to provide selection pressure towards improvement. While the true value of an individual is intended to be measured in terms of its effectiveness in the context of SVM-based classification, employing SVM classification every time an offspring needs to be evaluated is impractical from a computational cost perspective. Therefore, we employ a simpler fitness function that approximates the way in which spectrum features in an SVM are employed in the kernel function [30]. In essence, the fitness function rewards individuals that are over-represented in known HS sequences and under-represented in non-HS sequences. Given a k-mer w, the fitness value $f(w) = 100 * |c(w)_{\mathrm{HS}}/\texttt{total}_{\mathrm{HS}} - c(w)_{\mathrm{non-HS}}/\texttt{total}_{\mathrm{non-HS}}|$, where $c(w)$ counts the number of sequences containing w, and `total` normalizes by the number of known sequences in each class (HS or non-HS).

We remind the reader that the goal in this work is not only to employ these individuals as features for an improved classification, but also as motifs to shed light onto what sequence-based features are employed by HS regulatory sequences to recognize and interact with DNA-binding proteins. The chosen fitness function calculates the absolute difference between the percentage of known HS sequences and the percentage of non-HS sequences that contain an individual. An individual that is found in all HS sequences but no non-HS sequences, or alternatively in all non-HS sequences but no HS sequences, will have the highest score of 100. An individual found in the same percentage of non-HS sequences and HS sequences will have the lowest score of 0.

We have considered other functions, such as a quadratic fitness function that assigns high scores to individuals that are discriminative; that is, even individuals that are over-represented in non-HS sequences but under-represented in non-HS sequences. Low scores would be assigned to noisy individuals that are not discriminative of either HS or non-HS sequences. However, the goal of this work is to obtain insight into subsequences that are direct signals of HS regulatory sequences. This consideration and other potential extensions are discussed in section 4.

2.4.4 A Parallel Evolutionary Algorithm or Speciation

As described, our EA populations contain motifs of varying length. That raises a well-known issue when choosing two parents to produce an offspring via crossover: should the parents be of the same length? If so, crossover never produces offspring that have a different length than their parents. Since the specified mutation operator does not change the length either, there is no way to generate and maintain length diversity in the population. On the other hand, it is often the case in both nature and evolutionary algorithms that offsprings produced by structurally dissimilar parents are inviable.

In our case, preliminary experiments indicated better results when recombining parents of equal length. Since we were interested in exploring a limited set of motif lengths (6-12), we adopted an island-model approach in which each island contains individuals of the same length (i.e., one motif species), and evolves in isolation and in parallel with other

islands without migration. The best results, both in terms of the absolute quality of top-scoring motifs and the accuracy achieved by the SVM with these features, were obtained when employing this island model in our method. Our observations are in line with recent research that shows that an island model may yield more fit chromosomes rather than crossbreeding different species [40, 9].

2.5 Putting it All Together

The top-scoring individuals from each of the islands are collected (migrate) into a hall of fame registry. Out of this group, 25 top-scoring sequences are then designated as motifs and are supplied as features to the RBF kernel of an SVM trained on HS and non-HS sequences. Performance of an SVM that uses these motifs as features is compared with baseline SVM classifications that employ spectrum features.

3. RESULTS

Implementation Details.

The proposed method is implemented in Java and was run on an Intel Core2 Duo machine with 4GB RAM and 2.66GHz CPU. The EA implementation builds upon the Evolutionary Computation software that accompanies a textbook on the topic [9]. The method integrates genomic sequences with utilities for bioinformatics by making use of the open-source Biojava project [18]. Finally, our implementation of SVM in this work builds upon the LibSVM package [11].

3.1 Fitness Analysis of GA-obtained Motifs

Since it is not possible to determine *a priori* the motif length that would allow obtaining good motifs for recognition of HS sequences, we have conducted several independent experiments, where we vary the motif length from 6 to 12. Additionally, we compare results obtained from the island model (detailed in section 2) with results obtained from a variable-length EA, where individuals in the same population have variable lengths from 6 to 12. For each motif length selection, including variable-length and motif island model, 30 different runs of the same experiment are analyzed for consistency of obtained results. Fitness scores are first averaged over individuals in each generation and then averaged over 30 runs for each experiment. Figure 1 plots these average fitness scores every 100 generations.

Figure 1 shows the average fitness obtained over $5,000$ generations for each experiment. The highest fitness scores are obtained from the EA that employs the island model. Slight differences in the fitness scores are observed when using motifs of length 6, 7, 8, or 9. Moreover, average fitness scores obtained after $3,000$ generations are essentially similar for EAs that employ the island model or motifs of length 6, or 7. The results in Figure 1 shed additional insight into why the SVM-based classification in [30] obtains high classification accuracy when employing 6-mers as spectrum features in the kernel function.

Figure 1 also shows a sharp decline in fitness scores when increasing the motif length beyond 9, as indicated by the average fitness plotted for experiments that employs motifs of lengths 10, 11,12 or variable-length motifs in the same generation. These results suggest that the EAs with higher motif lengths do not sufficiently explore the solution space. Bigger population sizes are probably needed to enhance the exploration of the high-dimensional solution spaces that ensue when employing longer motifs.

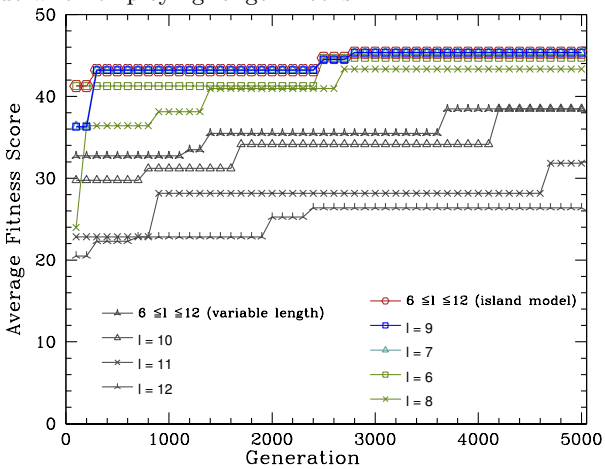

Figure 1: Fitness scores averaged first over individuals in a generation and then over 30 runs for each experiment are plotted every $1,000$ generations.

The highest average fitness scores shown in Figure 1 approach values of 48. According to the fitness function employed in this work (detailed in section 2, these scores mean that motifs in a generation (in the best-performing experiments) are on average in 48% more HS sequences than non-HS sequences. In order to show fitness scores of individuals generated in the course of a specific experiment, we plot in Figure 2 the fitness landscape associated with the solution space explored by an EA on motifs of length 7. The shown landscape higlights the fact that the solution space explored by the method contains a vast number of sub-optimal motifs and numerous local maxima (peaks in Figure 2).

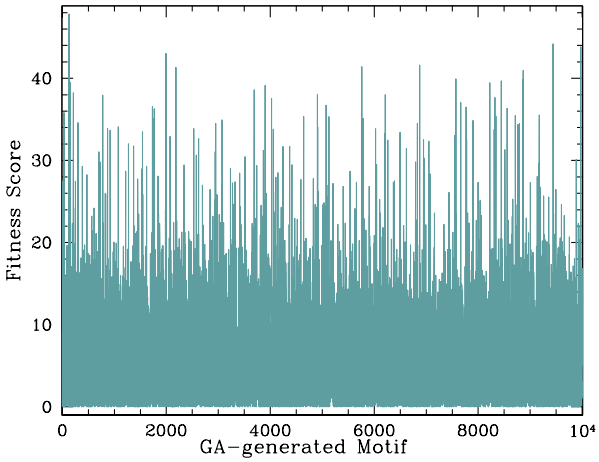

Figure 2: Fitness scores are drawn for each of the motifs generated during the course of an EA operating in a space of motifs of length 7.

Considering the results showcased in Figure 1, the top-scoring motifs obtained from the island model EA are employed as features for the following SVM-based classification.

3.2 Performance Analysis of SVM-based Classification Experiments

We conducted a systematic analysis that investigates whether employing the EA-obtained motifs as features improves the

classification accuracy in the context of an SVM with an RBF kernel. Employing all 6-mers as spectrum features and counting their frequency of occurrence in the RBF kernel is the first experiment that we conducted in order to obtain a baseline accuracy. Our additional experiments considered the following settings: EA-obtained motifs versus spectrum features in combination with EA-obtained motifs. Additionally, we varied the input to the RBF kernel: considering frequency of occurrence of a substring feature in an input sequence vs. noting only the presence or absence of a substring feature in an input sequence with a boolean value.

Experiment	$\mu \pm \sigma$	min	max
6-mers, frequency	77.27±0.00	77.27	77.27
motifs, frequency	84.41±2.70	79.11	88.12
motifs, boolean	83.47±2.92	78.12	89.11
6-mers+motifs, frequency	87.77±2.79	81.19	94.06
6-mers+motifs, boolean	85.40±3.53	78.12	90.10

Table 2: Average μ, standard deviation σ, minimum, and maximum values obtained from the 10-fold cross-validation are shown for each experimental setting. Values are in %.

Average accuracies obtained from a 10-fold cross-validation for each of the experimental settings listed above are summarized in Table 2. The spectrum,boolean setting is not conducted, as work in [30] shows that spectrum,frequency is needed to obtain good cross-validation accuracies. Table 2 shows that introduction of the EA-obtained motifs in the features employed by an SVM classification improves the average cross-validation accuracy by at least 5% over the baseline SVM classification that uses 6-mers as features. Additionally, the frequency of occurrence improves SVM performance over the boolean input to the RBF kernel.

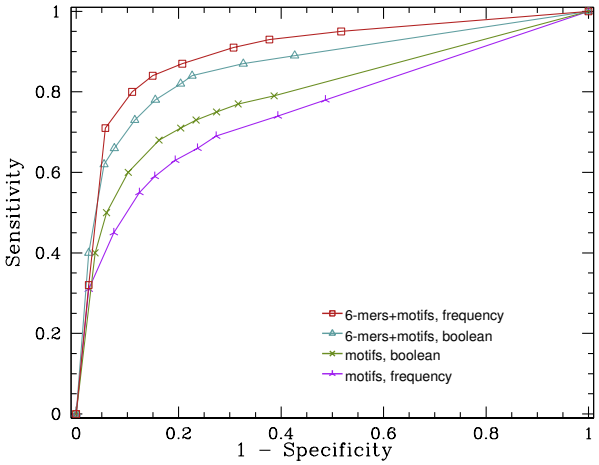

Figure 3: ROC curves for each experimental setting are shown in different colors, as listed in the legend. Sensitivity is plotted as a function of 1 - specificity.

The highest average cross-validation accuracy (88%), more than 10% higher than the baseline, was obtained when employing both EA-obtained motifs and 6-mers (and the frequency of occurrence in the RBF kernel) as features for the SVM. Interestingly, employing only the motifs improves upon the baseline by 7% but does not yield the highest overall accuracy. This points to the fact that potentially interesting 6-mers may not represented among the top-scoring motifs. Considering that EAs conduct a stochastic exploration of the solution space, it is possible that interesting local maxima may not be achieved as opposed to an exhaustive enumeration of 6-mers. We point out that the solution space considered by the EA is much larger than that of possible 6-mers, since we employ the IUPAC ambiguous code for DNA sequences instead of the 4-letter alphabet of DNA.

The ROC curves obtained for the experimental settings that achieve the highest cross-validation average accuracies are shown in Figure 3. Sensitivity is measured as TP/(TP + FN), and specificity is TN/(FP+TN), where TP, TN, FP, and FN refer to the number of true positives, true negatives, false positives, and false negatives. The results in Figure 3 agree with those highlighted in Table 2 (note that the average cross-validation accuracies in Table 2 correspond to the area under the ROC curves), showcasing that combining the EA-obtained motifs with 6-mers and employing frequency of occurrence in the RBF kernel yields the highest accuracy.

3.3 Analysis of Relationship between Fitness and Accuracy

The fitness in our EA-based method does not employ the SVM classifier but depends on occurrences of a motif in HS and non-HS sequences (detailed above). The following table tracks the relationship between the average fitness over the top-scoring motifs employed as features in the SVM classification and the resulting accuracy of the SVM classification.

Motif Length	l=6	l=9	l=12
Avg. Fitness	13.2 24.9 38.9	12.6 20.4 27.9	6.4 13.6 19.0
Accuracy	80.3 83.2 84.1	74.0 78.5 79.9	72.4 75.4 77.1

Table 3: Three experiments are run for each motif length $l \in \{6, 9, 12\}$. The average fitness over the top 25 motifs is shown versus the accuracy of the SVM classification that employs the evolved motifs.

The results in Table 3 show that there is a positive correlation between average fitness over the top-scoring motifs and the resulting accuracy of the SVM employing these motifs. This result further supports our employment of a simpler fitness function to evolve meaningful motifs while improving the accuracy of the SVM classification.

3.4 Analysis of Top-scoring Motifs

Classification is always more valuable if the features it identifies correspond to biologically-meaningful motifs [21, 20]. Therefore, we subject the top 25 motifs obtained by our EA algorithm to a multiple sequence alignment analysis. The alignment is carried out with clustalw 2.0 [23] to reveal any nucleotide patterns shared among the top-scoring motifs. The IUB scoring matrix, which specifies symbol matching scores and gap penalties, is employed to align sequences generated from the IUPAC code. Figure 4 shows the alignment that yields the highest overall additive score.

Inspection of the alignment in Figure 4 reveals that the CG pair is abundant and shared among all motifs. This agrees with analysis in [30], which shows that CG is abundant in HS sequences and confers high separation power to SVM. The aligment reveals additional patterns that go beyond dinucleotides. For instance, patterns like CGM, CGC, CGS, CGH, CGN, CSG, and even longer ones, such as CGMS, CGMSN, and CGSBN are well-preserved among the top-scoring motifs. The presence of ambiguous symbols

in our motifs allows revealing such patterns that possibly encode the needed chemical properties at the interaction sites between HS sequences and DNA-binding enzymes. A larger detailed biological study may draw important insight from this analysis into the binding motifs that facilitate the DNA-protein associations around the HS sites.

-	-	-	-	-	-	-	C	G	M	S	N	G	-	-
-	-	-	-	-	-	H	C	G	M	S	N	S	-	-
-	-	-	-	D	N	R	C	G	M	S	V	S	-	-
-	N	H	S	B	N	G	C	G	S	S	-	-	-	-
-	-	S	R	B	S	V	C	G	-	-	-	-	-	-
-	-	-	R	K	B	V	C	G	B	S	B	-	-	-
-	-	-	-	-	-	-	C	G	C	B	H	K	-	-
-	-	V	V	D	R	B	C	G	C	B	N	-	-	-
-	-	-	-	N	B	-	C	G	C	D	R	N	-	-
-	-	-	H	H	K	B	C	G	S	-	-	-	-	-
-	-	-	H	H	K	B	C	G	S	B	N	N	-	-
-	-	-	-	-	K	B	C	G	S	B	N	H	Y	-
-	-	-	-	-	-	-	C	G	B	B	S	C	-	-
-	-	-	-	-	B	S	C	G	S	K	S	-	-	-
-	-	-	B	V	G	B	C	G	D	-	-	-	-	-
-	-	-	-	-	V	K	C	G	H	N	H	V	-	-
N	H	-	-	-	V	R	C	G	H	K	B	-	-	-
-	-	-	-	-	V	B	C	G	N	Y	Y	V	-	-
S	H	N	Y	S	B	D	C	G	N	-	-	-	-	-
-	-	-	-	-	N	M	C	G	N	M	V	M	D	N
-	-	-	D	S	K	-	C	S	G	-	-	-	-	-
-	-	-	-	S	K	-	C	S	G	N	M	V	-	-
-	-	-	B	S	G	-	C	S	G	-	-	-	-	-
-	-	-	C	S	G	V	C	B	Y	N	-	-	-	-
-	-	-	N	Y	G	-	C	S	B	S	N	M	-	-

Figure 4: Highest-scoring alignmen of top-scoring motifs, as obtained through clustalw 2.0 [23]. Horizontal lines mark gaps in the alignment.

4. CONCLUSIONS

We have proposed an EA-based method to assist in mapping regulatory elements through recognition of HS sequences. The method operates under the hypothesis that HS sequences employ short motifs to interact with DNA-binding proteins. An EA-based search of both length and sequence space yields motifs of variable lengths that significantly improve recognition of HS sequences in the context of an SVM-based classifier. Compared with existing machine learning approaches that use standard k-mer spectrum features for recognition of HS sequences, the introduction of motifs as features improves classification accuracy by more than 10%.

Employing ambiguous symbols in the representation of motifs significantly expands the search space and in concert with the island model allows identifying high-quality motifs that discriminate between HS and non-HS sequences. A complete list of the motifs employed as features in the SVM classification conducted in our experimental analysis is available by request from the corresponding author.

Alignment of top-scoring motifs reveals shared patterns that, due to presence of ambiguous symbols, are not limited to subsequences of specific nucleotides but encode general chemical properties and possibly complex dependencies between positions. A detailed biological study may draw insight from anaysis of these patterns into the features that facilitate DNA-protein associations around HS sites.

As more evidence and data emerge from high-throughput experiments, automatic recognition of HS sequences may be further refined to account for possible associations with specific classes of regulatory elements. Incorporating a learning component in the fitness function employed by our EA may additionally allow improving both motif quality and classification accuracy. Advances in and combinations of experimental and computational techniques promise a comprehensive annotation of the entire human genome. Such a feat would improve our understanding of phenotypical differences and elucidate reasons for various regulatory diseases.

5. ACKNOWLEDGMENTS

We thank Rezarta Dogan for comments on this manuscript.

6. REFERENCES

[1] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. USA*, 98(2):757–762, 2002.

[2] M. Blanchette, A. R. Bataille, X. Chen, C. Poitras, J. Laganiere, C. Lefebvre, G. Deblois, V. Giguere, V. Ferretti, D. Bergeron, B. Coulombe, and F. Robert. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, 16(5):656–668, 2006.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152. ACM Press, 1992.

[4] B. Burgess-Beusse, C. Farrell, M. Gaszner, M. Litt, V. Mutskov, F. Recillas-Targa, M. Simpson, A. West, and G. Felsenfeld. The insulation of genes from external enhancers and silencing chromatin. *Proc. Natl. Acad. Sci. USA*, 99(S4):16433–16437, 2002.

[5] L. C., E. E., A. Cohen, A. Weston, and W. S. Noble. Mismatch string kernels for SVM protein classification. *Neur. Inf. Proc. Sys.*, 15(4):1441–1448, 2002.

[6] I. Committee. Nomenclature committee of the international union of biochemistry (nc-iub). nomenclature for incompletely specified bases in nucleic acid sequences. recommendations 1984. 229(2):75–88, 1985.

[7] Consortium IHGS. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[8] E. Davidson. *Genomic regulatory systems: Development and evolution*. Academic Press, New York, NY, 2001.

[9] K. A. De Jong. *Evolutionary computation: a unified approach*. MIT Press, Cambridge, MA, 2001.

[10] M. O. Dorschner, M. Hawrylycz, R. Humbert, J. C. Wallace, A. Shafer, J. Kawamoto, J. Mack, R. Hall, J. Goldy, P. J. Sabo, A. Kohli, Q. Li, M. McArthur, and J. A. Stamatoyannopoulos. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods*, 1(3):219–225, 2004.

[11] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using the second order information for training SVM. *J. Mach. Learn. Res.*, 6(1532-4435):1889–1918, 2005.

[12] G. Felsenfeld. Chromatin unfolds. *Cell*, 86(1):13–19, 1996.

[13] G. Felsenfeld. Controlling the double helix. *Nature*, 421(6921):448–453, 2003.

[14] D. S. Gross and W. T. Garrard. Nuclear hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, 57:159–197, 1988.

[15] T. Habib, C. Zhang, J. Y. Yang, M. Q. Yang, and Y. Deng. Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition. *BMC Genom.*, 9(Suppl1):S1–S16, 2008.

[16] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[17] D. R. Higgs, D. Vernimmen, J. Hughes, and R. Gibbons. Using genomics to study how chromatin influences gene expression. *Annu. Rev. Genom. Human Genet.*, 8:299–325, 2007.

[18] R. C. Holland, T. A. Down, M. Pocock, A. Prlic, D. Huen, K. James, S. Foisy, A. Draeger, A. Yates, M. Heuer, and M. J. Schreiber. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008.

[19] M. L. Howard and E. Davidson. cis-Regulatory control circuits in development. *Dev. Biol.*, 271(1):109–118, 2004.

[20] R. Islamaj-Dogan, L. Getoor, and W. J. Wilbur. A feature generation algorithm with applications to biological sequence classification. In H. Liu and H. Motoda, editors, *Computational Methods of Feature Selection*. Springer, Berling, Heiderlberg, 2007.

[21] R. Islamaj-Dogan, L. Getoor, W. J. Wilbur, and S. M. Mount. Features generated for computational splice-site prediction correspond to functional elements. *BMC Bioinformatics*, 8:410–416, 2007.

[22] A. Kernytsky and B. Rost. Using genetic algorithms to select most predictive protein features. *Proteins: Struct. Funct. Bioinf.*, 75(1):75–88, 2009.

[23] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustalw and clustalx version 2. *Bioinformatics*, 23(21):2947–2948, 2007.

[24] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: a string kernel for svm protein classification, 2002.

[25] C. H. Lowrey, D. M. Bodine, and A. W. Nienhuis. Mechanism of DNase I hypersensitive site formation within the human globin locus control region. *Proc. Natl. Acad. Sci. USA*, 89(3):1143–1147, 1992.

[26] M.-Z. Lu, C. L. P. Chen, and J.-B. Huo. Optimization of combined kernel function for svm by particle swarm optimization. In *IEEE Intl Conf on Machine Learning and Cybernetics*, volume 2, pages 1160–1166, Baoding, China, 2009.

[27] M.-Z. Lu, C. L. P. Chen, J.-B. Huo, and X. Wang. Optimization of combined kernel function for svm based on large margin learning theory. In *IEEE Intl Conf on Systems, Man and Cybernetics*, pages 353–358, Singapore, 2008.

[28] G. A. Maston, S. K. Evans, and M. R. Green. Transriptional regulatory elements in the human genome. *Annu. Rev. Genom. Human Genet.*, 7:29–59, 2006.

[29] W. S. Noble. Support vector machine applications in computational biology. In B. Schoelkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 2004.

[30] W. S. Noble, S. Kuehn, R. Thurman, M. Yu, and J. A. Stamatoyannopoulos. Predicting the in vivo signature of human gene regulatory sequences. *Bioinformatics*, 21(Suppl 1):i338–i343, 2005.

[31] S.-Y. Ohn, H.-N. Nguyen, D. S. Kim, and J. S. Park. Determining optimal decision model for support vector machine by genetic algorithm. In J. Zhang, J.-H. He, and F. Y., editors, *Lecture Notes in Computer Science: Computational and Information Science*, volume 3314, pages 895–902. Springer, 2005.

[32] A. Prakash, M. Blanchette, S. Sinha, and M. Tompa. Motif discovery in heterogeneous sequence data. In *Pac. Symp. Biocomput.*, pages 348–359, 2004.

[33] R. Reed and R. Mark. *Neural smithing: supervised learning in feed-forward artificial neurral networks*. MIT Press, Cambridge, MA, 1999.

[34] P. J. Sabo, R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Genome-wide identification of DNase I hypersensitive sites using active chromatin sequence libraries. *Proc. Natl. Acad. Sci. USA*, 101(13):4537–4542, 2004.

[35] S. Sinha and M. Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucl. Acids Res.*, 30(24):5549–5560, 2002.

[36] J. A. Stamatoyannopoulos, A. Goodwin, T. Joyce, and C. H. Lowrey. NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO J.*, 14(1):106–116, 1995.

[37] D. Talbot, J. N. J. Philipsen, P. Fraser, and F. Grosveld. Detailed analysis of the site 3 region of the human beta-globin dominant control region. *EMBO J.*, 9(7):2169–2177, 1990.

[38] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23(1):137–144, 2005.

[39] V. N. Vapnik. *Statistical learning theory*. Wiley & Sons, New York, NY, 1998.

[40] K. Vertanen. Genetic adventures in parallel: Towards a good island model under PVM, 1998.

[41] C. Wu. The $5'$ ends of drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, 286(5776):854–860, 1980.

[42] X. H. Zhang, K. A. Heller, I. Hefter, C. S. Leslie, and L. A. Chasin. Sequence information for the splicing of human pre-mrna identified by support vector machine classification. *Genome Res.*, 13(12):2637–2650, 2003.

[43] A. Zien, G. Raetsch, S. Mika, B. Schoelkopf, T. Lengauer, and K. R. Mueller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.