2014 IEEE Congress on Evolutionary Computation (CEC)
July 6-11, 2014, Beijing, China

GAMI-CRM: Using de novo motif inference to detect cis-regulatory modules

Jeffrey A. Thompson and Clare Bates Congdon
Department of Computer Science
University of Southern Maine
Portland, Maine 04104
Email: jeffrey.ahearn.thompson@maine.edu, congdon@usm.maine.edu

***Abstract*—In this work, we extend GAMI (Genetic Algorithms for Motif Inference), a *de novo* motif inference system, to find sets of motifs that may function as part of a cis-regulatory module (CRM) using a comparative genomics approach. Evidence suggests that most transcription factors binding sites are part of a CRM, so our new approach is expected to yield stronger candidates for *de novo* inference of candidate regulatory elements and their combinatorial regulation of genes. Thanks to our genetic algorithms based approach, we are able to search relatively large input sequences (100,000nt or longer). Most current computational approaches to identifying candidate CRMs depend on foreknowledge of the processes that the genes they regulate are involved in. In comparison with one leading method, Cluster-Buster, our prototype *de novo* approach, which we call GAMI-CRM, performed well, suggesting that GAMI-CRM will be particularly useful in predicting CRMs for genes whose interactions are poorly understood.**

I. Introduction

A cis-regulatory module (CRM) is a region of noncoding DNA that regulates the function of a gene. Specific transcription factor binding sites are grouped within the region and act together to affect gene expression [1]. This allows relatively few transcription factors to participate in complex differential expression patterns of a larger number of genes [2]. Therefore, knowledge of CRMs is an important part of understanding gene regulatory networks and disease.

In this work, we build on a motif inference system, developed in prior work, called GAMI (Genetic Algorithms for Motif Inference) [3], [4]. GAMI is used for *de novo* identification of candidate regulatory elements in noncoding DNA. GAMI uses a genetic algorithms (GA) search to identify patterns (motifs) that recur in multiple sequences that are being compared. These sequences are noncoding DNA upstream from orthologous genes in divergent species. Conservation among species with common ancestors is often used as an indicator of possible functional regions, since such regions should be under selective constraint.

In this paper, we demonstrate how our new prototype tool, GAMI-CRM, discovers clusters of motifs that can be used to predict CRMs. In the remainder of this paper, Section II explains relevant background, including characteristics of CRMs and computational approaches. Section III describes the system design of GAMI-CRM. Section IV explains our research methodology, including the data curated for this work and the experimental design. Section V presents the results; Section VI discusses the implications of the results, and Section VII describes future work.

II. Background

A CRM acts as a unit in regulating the function of a gene [5], and it is therefore desirable to identify a CRM as a whole, in addition to the individual binding sites within it. However, successful prediction requires that CRMs have features that distinguish them from surrounding, non-regulatory, sequence. Despite the importance of CRMs, their characteristics are not yet well understood. The task of distinguishing noncoding regulatory DNA from non-regulatory DNA is challenging, from either a computational or biochemical perspective. Until more CRMs are validated, knowing their common properties is difficult, but without accurate predictions it is challenging to validate more CRMs.

The largest database of validated CRMs is RedFly [1], which contains hundreds of validated *Drosophila* CRMs. A study of confirmed CRMs in the RedFly database [6] found certain characteristics in common among many of them. These include elevated GC content, conservation, and dense clustering of TFBSs. Other work has shown that combinations of biochemical markers are strong indicators of CRMs [7].

A. Approaches to Identifying CRMs

There have been numerous attempts to develop methods of identifying CRMs, as reviewed in [8], [2], [7], in addition to our own approach [9]. Generally speaking, these approaches look for conservation of noncoding DNA, clustering of known transcription factor binding sites (TFBSs), or biochemical markers (such as histone modification) [7]. Although each of these approaches has merit, they each have limitations.

Known CRMs tend to be more highly conserved across related species than other noncoding DNA [2], [6]. Within a CRM, the individual TFBSs show even higher conservation [2]. However, CRMs vary considerably in length. Most CRM predictors that use conservation examine a set of multiple alignments using a fixed window size for the CRM. With this approach, CRMs with fewer binding sites may cause a window to have seemingly low overall conservation, when in reality the window is too large. This may help explain why such alignments fail to detect a number of CRMs [7]. An additional limitation of alignments is that they are unable

978-1-4799-1488-3/14/$31.00 ©2014 IEEE

to consider CRMs in which the binding sites may occur in different orders [8].

Although individual binding sites can occur at random, this is significantly less likely to happen for clusters of binding sites [10], making such clusters a potentially useful filter for detecting CRMs. However, most systems that look for clusters of TFBSs are capable of searching only for known sites. These sites are frequently described as position weight matrices (PWMs), which are compiled by comparing known sites from multiple DNA sequences (often across species). They are essentially a table of how frequently a nucleotide occurs at each location in a transcription factor binding site throughout the sequences. Therefore, to use a system that scans for known TFBSs, one must obtain PWMs that describe the sites to search for. This in turn means that one must have some understanding, in advance, of how a gene is regulated to use this technique. Although large databases of PWMs, such as TRANSFAC [11], do exist, loading all available PWMs into a system can lead to many false positives and potentially interfere with the discovery of the true CRMs [12]. However, if the number of PWMs used is too small, CRMs containing some unknown sites will appear to have less dense clustering and may not be detected. Currently, this approach has shown limited success [2].

Biochemical prediction is based on using epigenetic features such as histone modification marks and CHiP-seq data to predict CRMs. When data is available, this can be highly effective [7]. However, CHiP-seq depends on antibodies being available for the transcription factors in question, which is often not the case. Additionally, proteins frequently bind only in certain tissues or at specific developmental stages, which may require numerous experiments depending on the research question. This means the time and cost associated with this approach may be prohibitive unless the data are already available. Furthermore, some criticism has emerged that binding of a transcription factor in itself does not show regulation of a gene and that functional elements should be conserved as well [13].

B. Cluster-Buster

A full review of CRM prediction tools is beyond the scope of this article. Above, we have outlined some of the general principles and challenges of prediction. Here, we discuss a method that is representative of the approach common to nearly all tools: scanning sequences against a library of PWMs.

Cluster-Buster is a single sequence approach [14]. It follows Cister [15] and COMET (Clusters of Motifs E-value Tool) [16], which were developed by the same research group. As its name implies, it searches for clusters of transcription factor binding sites, attempting to find regions that can be statistically differentiated from the background sequence using a hidden Markov model (HMM). Cluster-Buster uses a heuristic approach that completes in time linear with the length of the input, rather than searching for every possible cluster. Despite the fact that its last update was in 2007, Cluster-Buster remains a competitive choice. In a review in 2010 [8], Cluster-Buster was one of the top performers, despite the fact that it does not use conservation as part of its prediction.

C. De Novo CRM Inference

A few other *de novo* CRM inference systems do exist. Examples include an evolutionary Monte Carlo method known as EMCModule [17], a Gibbs sampling method [18], and a Bayesian approach known as Cis-Module [19], in addition to our own approach, GAMMI [9]. Each of these projects has certain limitations that would be helpful to overcome. EMCModule does not infer the binding sites themselves, these are provided as input to the algorithm after running a *de novo* motif inference tool, as well as downloading PWMs from a database. However, EMCModule suffers from poor specificity when more than about 100 motifs are searched for. Therefore, a choice must be made before running to tool as to which factors may be involved in regulation [17]. The Gibbs sampling approach of Thompson et al. is designed specifically to be used on input sequences that are likely to be co-regulated [18]; therefore, gene expression analysis is a necessary first step. Cis-Module requires parameters to be set for the likely length of the CRM and the number of TFs involved [19]. Although guidance is provided on how to calculate these parameters, they contribute significantly to the complexity of Cis-Module. Finally, our own prior work, GAMMI, relied on two evolutionary computation steps. Our goal is to develop an approach to *de novo* CRM prediction that can search for an unlimited variety of motifs, in a CRM of unknown size, comprised of an unknown number of TFs, using the DNA sequence alone.

D. The GAMI Algorithm

The target of GAMI's search is an N-mer that appears at least once in each input sequence. However, we allow imperfect matches, so a motif does not need to be fully represented in a sequence. Instead, N-mers that match more strongly are considered stronger motifs. The N-mer itself is a sequence of N bases from the set {`A, C, G, T`}. For example, if we are search for 8-mers, possible motifs identified include `CATGCAAT`, `TAGGAACT`, `ACTTACGT`, etc.

Aside from exhaustive search, there is not an algorithmic way to calculate the best motifs for a set of sequences, and depending on the number of sequences being examined, the sequence length, and the motif length, exhaustive search can be prohibitively computationally expensive. Therefore, most approaches to motif inference use some sort of heuristic search technique; GAMI uses a GA search. GAMI searches for motifs on the {`A, C, G, T`} alphabet as described above; the fitness function is a linear measure reflecting the quality of the match across the input sequences [3], [4].

Although GAMI implements a relatively standard genetic algorithm, it uses a high level of elitism (50%). Therefore, half the population of candidate solutions is carried over from one generation to the next. This means that when GAMI finishes running, we are left with hundreds of candidate solutions in the form of motifs that are ranked by strength of conservation.

An important feature of GAMI is that it does not depend on multiple alignment to find conserved motifs. Motifs are evolved and ranked for conservation by the GA. Therefore, GAMI is capable of finding conserved elements even within noncoding regions that are widely diverged.

III. System Design

GAMI-CRM is based on GAMI, and leverages the fact that CRMs (as well as the TFBSs within them) tend to be more highly conserved than surrounding noncoding DNA.

As a first step, we use GAMI to look for motifs that are somewhat longer than typical TFBSs. Here we use motifs that are 20nt in length, although it is likely that other settings would also perform well. As mentioned previously, GAMI will output a ranked list of hundreds of candidate solutions. GAMI-CRM then maps each motif onto the query sequence (the particular sequence we are interested in finding CRMs for); in many cases, these motifs will form overlapping regions. The best conserved sets of overlapping solutions should be more likely to occur within CRMs, because they will describe a region that exhibits conserved motifs in close proximity. Additionally, in [20] it was found that Drosophila enhancers are enriched in ungapped conserved blocks 20nt or more in length.

After this step, GAMI-CRM will have a list of candidate CRMs of varying length. The next step is to rank the candidates by their overall conservation. Although the individual motifs have a conservation score, they are relatively short, ungapped sequences. However, within the CRM, there is a strong likelihood of gaps, especially between transcription factor binding sites. For this prototype, GAMI-CRM uses BLAST (Basic Local Alignment Search Tool) [21] to perform this step. A BLASTN (nucleotide BLAST) query is performed (version BLASTN 2.2.28+), which aligns the candidate CRMs from the target species back to the input sequences given to GAMI (excluding the target species). The score for each candidate CRM is then taken as the sum of BLAST scores for all significant hits.

Finally, GAMI can be run on the same input sequences to look for shorter motifs, perhaps 10nt, and we can find the intersection of these short motifs with our predicted CRMs in order to predict the individual binding sites within. This makes the system a combined *de novo* motif and CRM inference tool.

Briefly then, our method involves four steps:

1) Run GAMI using a set of noncoding DNA sequences upstream of orthologous genes as input to look for motifs that are 20nt long. Although it is somewhat longer than a typical binding site, this length should help GAMI-CRM find overlapping sites.
2) Find overlapping solutions and combine them into longer subsequences.
3) Run BLASTN using the subsequences identified in the previous step to score their conservation against the set of input sequences.
4) Optionally, run GAMI again on the input data, to look for motifs that are 10nt long. This may help identify individual TFBSs within the candidate CRM.

At the end of this process we are left with a ranked set of candidate CRMs that should avoid some of the issues faced by other approaches. The conservation considered by GAMI is closer to the level of the TFBS, rather than the entire CRM. Tools that use fixed window sizes to examine alignments can calculate mistakenly low conservation in CRMs, or low density of sites, when forced to consider the entire window. Our proposed solution sidesteps this issue by considering conservation and clustering in a more dynamic fashion. We may predict a couple of smaller CRMs that are in fact part of a larger CRM, but our system should have better sensitivity for such cases. It is also worth considering that the reverse can also be a problem for systems with a fixed window size; they may predict a single CRM that is in fact multiple CRMs. This is currently a challenging issue in CRM prediction.

IV. Methodology

In prior work, we established that GAMI is an effective approach to inferring candidate regulatory elements such as TFBSs [3], [4]. Therefore, in this work, we focus on GAMI-CRM, and evaluate its performance as a prototype cis-regulatory module inference system. To that end, we ran GAMI-CRM on two benchmark datasets that are described below and compared its performance to a well established CRM inference system known as Cluster-Buster [14]. As mentioned previously, Cluster-Buster is an HMM based approach to predicting CRMs using PWMs that has been shown to perform well on a variety of data [8], [22]. In [8], there are two approaches that outperform Cluster-Buster on the data used in their evaluation, notably [23] and [24]. However, Cluster-Buster is widely used because of its robustness in a variety of data and its ease of use. Therefore, it is an appropriate choice in helping us to understand the performance of our prototype tool. We used the web interface to Cluster-Buster available at http://zlab.bu.edu/cluster-buster/, which was released in 2007.

To evaluate the performance of the two methods (GAMI-CRM and Cluster-Buster), we used a number of standard performance measures for classification tasks as described in [25]. These are based on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The measures are:

- Accuracy: (TP + TN) / (TP + FP + FN + TN)
- Sensitivity: TP / (TP + FN)
- Specificity: TN / (FP + TN)
- Positive Predictive Value: TP / (TP + FP)
- Negative Predictive Value: TN / (FN + TN)

For our evaluation we will define TP as the number of nucleotides predicted to be in a CRM and that are present in a validated CRM, TN as the number of nucleotides in the input that are not in a validated CRM and not in a predicted CRM, FP as the number of nucleotides predicted to be in a CRM but not found in a validated CRM, and FN as the number of nucleotides in a validated CRM but not predicted to be in a CRM. Only Accuracy uses all four values (TP, TN, FP, FN). Therefore, it usually provides the best indication of overall performance. However, there is no way of knowing

what the actual number of true positives is (there is no gene for which all regulatory elements are known to be annotated). Although we believe there is value in this analysis regardless, this limitation should be kept in mind.

A. Data

We curated two data sets for this preliminary work: the noncoding sequences upstream from ADSSL1 and SMYD1. In both cases, we used the entire sequence upstream of the target gene to the next known gene. For ADSSL1, the human sequence was 4586nt in length. The intergenic region upstream of SMYD1 in humans is 12133nt in length. Both of these genes play a role in the differentiation of skeletal muscle tissue [2]. They each contain a functional CRM that was validated in [2]. In that work, the CRM is identified in coordinates relative to the hg18 assembly of the human genome available through the UCSC Genome Browser [26]. Therefore, we curated the two human noncoding regions from the UCSC Genome Browser, so that we could identify the sequence of the validated CRMs. Since GAMI is a comparative genomics system, we also obtained sequences from all available species through NCBI's Entrez Gene [27]. The orthologs were identified by annotation for the above listed genes in the database. This resulted in 13 sequences in each data set.

B. GAMI Parameter Settings

For all experiments reported here, we used a population size of 1,000, a crossover rate of 0.8, and a mutation rate of 0.02. The number of trials was set at 200,000 (which refers to the number of fitness function evaluations; due to elitism and the ability to recognize when a reproduction operator has no effect, there is not a clean mapping between the number of trials and the number of generations). Fifty percent elitism was used to preserve the best 500 motifs in the population every generation. Therefore, at most 500 new motifs are created every generation, and the result of a run can be considered the 500 best solutions in the final population. The 80 percent crossover rate means that 80 percent of the remaining motifs are candidates for crossover (a total of 400). The 2 percent mutation rate means that a nucleotide in a solution has a 2 percent chance of being set to a random value (possibly the same as it was before). Rank-based selection was used. The motif length was set to 20. These settings are the default settings we generally use with GAMI.

C. Cluster-Buster Parameter Settings

For Cluster-Buster, we relied on the default settings. The Gap Parameter was set to 35 (the expected average distance between motifs). The Cluster Score Threshold was set to 5 (this determines which results will be reported). The Motif Score Threshold was set to 6 (similar to Cluster Score Threshold, but for motifs). Residue Abundance Range was set to 100 (this determines how far to look around a site to determine the relative frequency of each nucleotide). Pseudocount was set to 0.375 (which is an adjustment to the counts in the PWM that helps balance PWMs with fewer sequences). We did not filter lower complexity regions with Dust.

D. Experimental Design

For each dataset, GAMI was run twenty times and the best results from all runs were combined. This is our standard method of running GAMI to account for the stochastic nature of genetic algorithms. GAMI-CRM was run on the full dataset of noncoding sequences described above, while Cluster-Buster was run on the human sequence only. We compared the top three candidate CRMs from our results to those from Cluster-Buster. Cluster-Buster never reported more than three results with these data and settings, but our method tends to report more candidate CRMs.

As mentioned previously, Cluster-Buster requires the additional input of position weight matrices. For each dataset, we ran it in two ways and compared the results to our method. The first approach was simply to select the 16 PWMs that are provided with Cluster-Buster by default. These are: TATA, Sp1, CRE, ERE, NF-1, E2F, Mef-2, Myf, CCAAT, AP-1, Ets, Myc, GATA, LSF, SRF, Tef. We did not expect that this approach would work particularly well, because it has been observed that using too many PWMs can cause enough false positives to interfere with accurate identification [12]. The second approach was to use PWMs for TFs known to play a role in muscle differentiation. These were: Sp1, Mef-2, Myf, SRF, and TEAD as identified in [2]. Sp1, Mef-2, Myf, and SRF were already available from Cluster-Buster. The PWM for TEAD was retrieved from the JASPAR database of transcription factor PWMs [28].

GAMI-CRM does not use a hard-wired cutoff to reduce the number of candidate solutions; this is a user parameter. Candidates are scored and ranked and the user is left to decide how many of the solutions merit further investigation. Cluster-Buster produces far fewer predictions. This makes is difficult to compare its results directly to Cluster-Buster. However, we will assume that the validated CRM in each data set is the easiest to identify (and should therefore receive the highest scores). Therefore, we compared the number of results from each system that is equal to the highest number of clusters identified by one of the Cluster-Buster runs. For example, if Cluster-Buster with the muscle transcription factors identified 3 clusters and Cluster-Buster with the default PWMs identified 2 clusters, we will compare these results to the top 3 GAMI results.

V. Results

In this work, we have developed a prototype approach to extending GAMI to predict cis-regulatory modules in noncoding DNA, called GAMI-CRM. This approach uses conservation of noncoding regions upstream of orthologous genes and adjacency of conserved elements to make its predictions, and entails post processing the GAMI motifs to identify the candidate CRMs.

A. Benchmark Results

For the ADSSL1 data set, Cluster-Buster was run twice, once with the set of default PWMs and once with the set of PWMs of transcription factors known to play a role in muscle

System	Best Location (from TSS)	Length	in CRM
GAMI-CRM	-2050..-1951	99	yes
Cluster-Buster w/ Muscle PWMs	-2639..-1935	704	partly
Cluster-Buster w/ Default PWMs	-2016..-1906	110	yes

TABLE I
ADSSL1 RESULTS - GAMI-CRM VS. CLUSTER-BUSTER

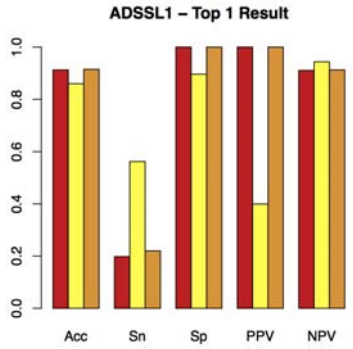

Fig. 1. Results for GAMI-CRM (red), Cluster-Buster with Muscle PWMs (yellow), and Cluster-Buster with Default PWMs (orange): Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Positive Predictive Value (PPV), and Negative Predictive Value (NPV)

System	Acc	Sn	Sp	PPV	NPV
GAMI-CRM	0.91	0.20	1.00	1.00	0.91
Cluster-Buster w/ Muscle PWMs	0.86	0.56	0.90	0.40	0.94
Cluster-Buster w/ Default PWMs	0.91	0.22	1.00	1.00	0.91

TABLE II
ADSSL1 CRM RECOVERY - GAMI-CRM VS. CLUSTER-BUSTER

differentiation. Cluster-Buster with or without muscle PWMs both identified a single cluster, so we compared them to the single top result obtained from GAMI-CRM. These results are shown in Table I and Figures 1 and 3.

For the SMYD1 data set, again Cluster-Buster was run twice. Cluster-Buster with the muscle PWMs identified 3 clusters and with the default PWMs identified 2 clusters. Therefore, the top 3 results from GAMI-CRM were used in the comparison shown in Table III and Figures 2 and 4. The accuracy and other results for each method are additionally shown in Table II and Table IV.

VI. DISCUSSION

A few things are immediately apparent from these results:

- With these data, both GAMI-CRM and Cluster-Buster were able to identify the validated CRM with a high degree of accuracy.
- The muscle PWMs enabled Cluster-Buster to predict the validated CRMs with greater sensitivity than either GAMI-CRM or Cluster-Buster with the Default PWMs.
- The accuracy and specificity of GAMI-CRM was greater with these data.
- Cluster-Buster with default PWMs was unable to identify the validated CRM in the SMYD1 data.

System	Location (from TSS)	Length	in CRM
Best Result			
GAMI-CRM	-442..-389	53	yes
Cluster-Buster w/ Muscle PWMs	-2692..-2522	170	no
Cluster-Buster w/ Default PWMs	-2696..-2518	178	no
2nd Best Result			
GAMI-CRM	-530..-479	51	yes
Cluster-Buster w/ Muscle PWMs	-8211..-8029	182	no
Cluster-Buster w/ Default PWMs	-8211..-7988	223	no
3rd Best Result			
GAMI-CRM	-7501..-7444	57	no
Cluster-Buster w/ Muscle PWMs	-587..-91	496	partly
Cluster-Buster w/ Default PWMs	–	0	no

TABLE III
SMYD1 RESULTS - GAMI-CRM VS. CLUSTER-BUSTER

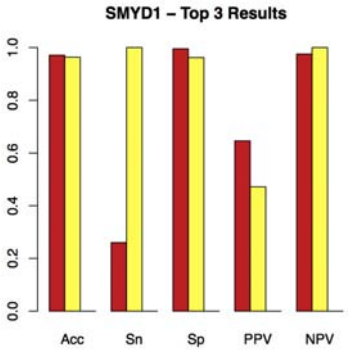

Fig. 2. Results for GAMI-CRM (red), Cluster-Buster with Muscle PWMs (yellow), and Cluster-Buster with Default PWMs (orange): Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Positive Predictive Value (PPV), and Negative Predictive Value (NPV)

System	Acc	Sn	Sp	PPV	NPV
GAMI-CRM	0.97	0.26	1.00	0.65	0.98
Cluster-Buster w/ Muscle PWMs	0.96	1.00	0.96	0.47	1.00
Cluster-Buster w/ Default PWMs	–	–	–	–	–

TABLE IV
SMYD1 CRM RECOVERY - GAMI-CRM VS. CLUSTER-BUSTER

As we note above, the muscle PWMs improved the sensitivity, but not the specificity, of Cluster-Buster on these data. However, by looking at Figure 1 and Figure 2 we can see that the increased sensitivity comes at a price. Cluster-Buster with muscle PWMs has a much higher false positive rate (reflected in its lower PPV score). This is particularly apparent in Figure 3. The green bar represents the CRM identified by Cluster-Buster, and the brown bars represent the individual motifs identified by Cluster-Buster. The blue and red bars represent the results from GAMI-CRM. By comparing these regions to the black bar labeled "ValidatedCRM", it is obvious that Cluster-Buster identified a region that extends well beyond the validated CRM. The region may include multiple CRMs, which may also have been identified by GAMI-CRM, but we are only considering the top result. The same effect can be

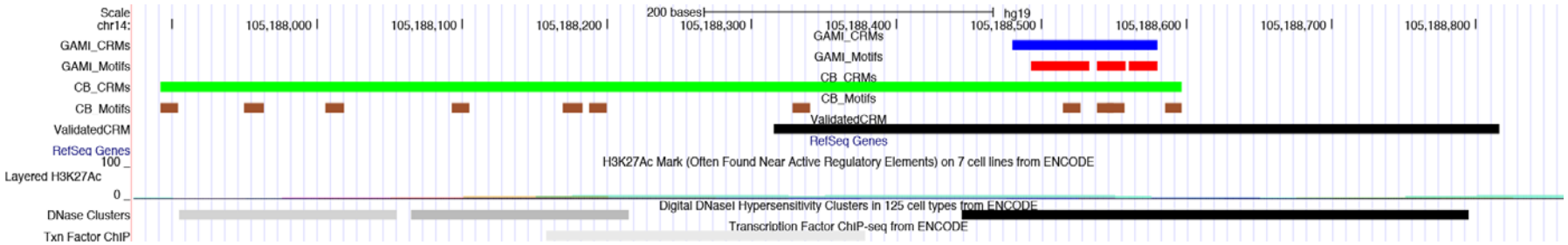

Fig. 3. Results of the GAMI-CRM and Cluster-Buster with muscle PWM runs on the ADSSL1 data visualized in tracks of the UCSC Genome Browser. The view is zoomed in to show approximately 1000 nucleotides. The blue and red bars are the results from GAMI-CRM and the green and brown bars are the results from Cluster-Buster. The bottom three tracks show epigenetic data from the ENCODE project that relates to regulatory regions [29]. Third from the bottom is the H3K27Ac histone modification mark, which is frequently found near regulatory activity. Second from the bottom are DNase hypersensitive clusters, which are regions of accessible chromatin often bound by transcription factors. The bottom track shows regions that were found bound by transcription factors measured by CHiP-Seq.

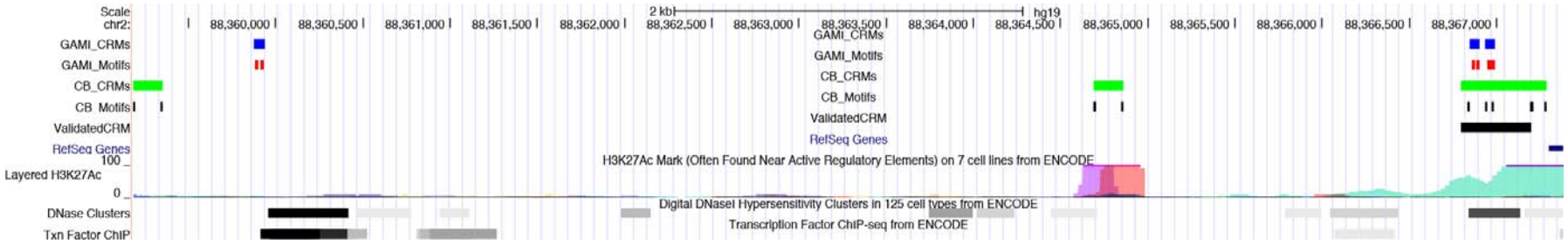

Fig. 4. Results of the GAMI-CRM and Cluster-Buster with muscle PWM runs on the SMYD1 data visualized using the UCSC Genome Browser. The view is zoomed in to show approximately 8000 nucleotides. The blue and red bars are the results from GAMI-CRM and the green and brown bars are the results from Cluster-Buster. The bottom three tracks show epigenetic data from the ENCODE project that relates to regulatory regions [29]. Third from the bottom is the H3K27Ac histone modification mark, which is frequently found near regulatory activity. Second from the bottom are DNase hypersensitive clusters, which are regions of accessible chromatin often bound by transcription factors. The bottom track shows regions that were found bound by transcription factors measured by CHiP-Seq.

seen in Figure 4 to a lesser extent, where the region identified by Cluster-Buster extends beyond the validated CRM.

We compared GAMI-CRM's single best result against Cluster-Buster's best result on the ADSSL1 data, because Cluster-Buster only identified a single cluster. However, if we had included the top two results from GAMI, its sensitivity would have increased to 0.37 with no drop in its specificity. This would additionally have increased its accuracy to 0.93.

Clearly, the issue of how many of the top results to consider for further analysis is not only difficult to decide but has a significant impact on the results. However, it also raises another issue. By examining Table III we can see that the top two results are both included within the validated CRM. They are also very close to each other. Currently our prototype has no means of connecting these solutions, but this may be possible by making the method slightly more sophisticated. As we mentioned earlier, the conservation of the CRM as a whole is often elevated in comparison to surrounding regions, as well as the GC content. Possibly we can take these into account when creating candidate solutions. Had we connected these regions together, the sensitivity of the SMYD1 results would have increased to .35, with no drop in specificity. If we had done the same for the top two results in the ADSSL1 data, the sensitivity would have been 0.62, with no drop in specificity. This would have increased the accuracy to 0.96. This just goes to show that these sorts of figures are quite dependent on how we define our solutions.

Another factor to consider is that there is only a single validated CRM for each gene, however, genes are frequently regulated by multiple CRMs. One way to think about this is to include epigenetic data in the analysis. In Figure 3 and Figure 4 we have visualized the results in the UCSC Genome Browser. The bottom three tracks of each image show a selection of epigenetic data that could be used as part of the analysis (there is more available). Third from the bottom is a track showing the H3K27Ac histone modification mark, which is often found in conjunction with active regulation. Second from the bottom is a track displaying DNase hypersensitive site (DHS) clusters, which are region of accessible chromatin within which transcription factors are often bound. Finally, the bottom track shows regions where transcription factors have been found to be bound through CHiP-Seq assay. It is worth keeping in mind that these data are measured within certain cell lines, sometimes a fairly limited set. They do not provide a complete picture of biochemical activity. Nevertheless, some interesting observation can be made in relation to our results. In Figure 3, the validated CRM overlaps only a single cluster of DHSs. Furthermore, that cluster is active in more cell lines than the other nearby clusters. Although the Cluster-Buster CRM overlaps all three DHS clusters in the figure, this may lend weight to the idea that multiple CRMs are present in the region. In Figure 4, CRMs were predicted by both GAMI-CRM and Cluster-Buster far upstream of the validated CRM (to the left of the figure). GAMI-CRMs prediction in this area overlaps a region that was shown to be bound by transcription factors in some cell lines (the bottom track). This may imply that the CRM is indeed functional. Similarly, the CRM predicted by Cluster-Buster in the middle of the figure overlaps a region with the H3K27Ac mark, which may lead one to consider that candidate as being plausible. Interestingly, in both data sets, the strongest epigenetic data is associated with the validated CRM (and the predictions of both systems).

It is probably no surprise that both GAMI-CRM and Cluster-Buster identified the validated CRMs with such a high degree of accuracy. The regions were originally identified for study using a variety of measures including conservation and clustering of transcription factor binding sites [2]. Nevertheless, our results suggest that both GAMI-CRM and Cluster-Buster are able to accurately identify CRMs when they exhibit these characteristics. However, the fact that Cluster-Buster failed to identify the validated CRM in the SMYD1 dataset when using the default PWMs shows an important difference between the two systems. GAMI-CRM is a *de novo* CRM predictor. In these experiments, GAMI-CRM had no need to know processes these genes were involved in in order to successfully identify validated CRMs that regulate them. This is not true of Cluster-Buster. One must have an understanding of how a gene might be regulated, in order to successfully predict CRMs using systems that depend on scanning with PWMs. The default PWMs included 4 of the 5 muscle-related PWMs, but even so, Cluster-Buster failed to identify the known CRM in the SMYD1 data. In addition, the PWMs must be available and a reasonably good fit for the transcription factor binding sites in question. When examining poorly understood genes, or particular tissues for which interaction data is unavailable, this is likely to be a significant limitation.

GAMI-CRM identified both known CRMs in these data with high accuracy. Thanks to its de novo inference, it did this without prior knowledge of the processes the CRM is involved in. Although Cluster-Buster performed well, the correct combination of matrices is necessary for accurate identification. Our results suggest that GAMI-CRM will be particularly useful in predicting CRMs for genes whose interactions are poorly understood.

VII. Future Work

The prototype we developed in this work performed well on the benchmark data sets, but we must validate on a greater number of data sets with known CRMs. Additionally, since genes are frequently regulated by more than one CRM, we will investigate GAMI-CRM's ability to detect such CRMs. Additionally, we will compare our approach to additional extant methods.

One limitation of our current approach is GAMI-CRM's reliance on BLAST for ranking the candidate CRMs. Although BLAST performs well, the use of alignment means that a candidate CRM may be scored low if the sites within it are in a different order than those in other sequences. However, this does not necessarily mean the CRM is not conserved. We will investigate the ramifications of this approach and possible alternatives in the future.

In previous work [9], we developed GAMMI, an evolutionary-computation approach to construct CRMs from a library of candidate motifs. This work was evaluated on artificial datasets. While the candidate motifs can come from GAMI or another motif inference system (and do not need to be verified PWMs), the system required a fixed window size. A stronger system can be expected to be realized by integrating these two approaches, using the candidate CRM window identification process described here to identify the windows, and GAMMI to identify the CRMs within these windows. We expect that this will make the use of BLAST tools unnecessary in this process. Additionally, GAMMI does not require the motif sites to be in the same order in the CRMs in different contexts, so this approach can be expected to have greater flexibility.

VIII. Acknowledgments

This project was supported by grants from the National Center for Research Resources (5 P20 RR024475-02) and the National Institute of General Medical Sciences (8 P20 GM103534-02) from the National Institutes of Health, a National Science Foundation (NSF) CAREER award (#953495), and NSF Cooperative Agreement No. HRD-0833567.

References

[1] M. S. Halfon, S. M. Gallo, and C. M. Bergman, "REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D594–598, Jan 2008.

[2] A. T. Kwon, A. Y. Chou, D. J. Arenillas, and W. W. Wasserman, "Validation of skeletal muscle cis-regulatory module predictions reveals nucleotide composition bias in functional enhancers," *PLoS Comput. Biol.*, vol. 7, no. 12, p. e1002256, Dec 2011.

[3] C. B. Congdon, C. Fizer, N. W. Smith, H. R. Gaskins, J. C. Aman, G. M. Nava, and C. J. Mattingly, "Preliminary results for gami: A genetic algorithms approach to motif inference." in *CIBCB'05*, 2005, pp. 97–104.

[4] C. B. Congdon, J. C. Aman, G. M. Nava, H. R. Gaskins, and C. J. Mattingly, "An evaluation of information content as a metric for the inference of putative conserved noncoding regions in DNA sequences using a genetic algorithms approach," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 5, pp. 1–14, 2008.

[5] D. M. Jeziorska, K. W. Jordan, and K. W. Vance, "A systems biology approach to understanding cis-regulatory module function," *Semin. Cell Dev. Biol.*, vol. 20, no. 7, pp. 856–862, Sep 2009.

[6] L. Li, Q. Zhu, X. He, S. Sinha, and M. S. Halfon, "Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses," *Genome Biol.*, vol. 8, no. 6, p. R101, 2007.

[7] R. C. Hardison and J. Taylor, "Genomic approaches towards finding cis-regulatory modules in animals," *Nat. Rev. Genet.*, vol. 13, no. 7, pp. 469–483, Jul 2012.

[8] J. Su, S. A. Teichmann, and T. A. Down, "Assessing computational methods of cis-regulatory module prediction," *PLoS Comput. Biol.*, vol. 6, no. 12, p. e1001020, 2010.

[9] D. J. Gagne and C. B. Congdon, "Preliminary results for GAMMI: Genetic algorithms for motif-module inference," in *Proceedings of the 2012 IEEE Congress on Evolutionary Computation*, X. Li, Ed., Brisbane, Australia, 10-15 June 2012, pp. 1309–1316.

[10] A. Wagner, "Genes regulated cooperatively by one or more transcription factors and their indentification in whole eukaryotic genomes." *Bioinformatics*, vol. 15, no. 10, pp. 776–784, 1999.

[11] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res.*, vol. 31, pp. 374–378, Jan 2003.

[12] K. Klepper, G. K. Sandve, O. Abul, J. Johansen, and F. Drablos, "Assessment of composite motif discovery methods," *BMC Bioinformatics*, vol. 9, p. 123, 2008.

[13] D. Graur, Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, and E. Elhaik, "On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE," *Genome Biology and Evolution*, vol. 5, no. 3, pp. 578–590, 2013. [Online]. Available: http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evt028

[14] M. C. Frith, M. C. Li, and Z. Weng, "Cluster-Buster: Finding dense clusters of motifs in DNA sequences," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3666–3668, Jul 2003.

[15] M. C. Frith, U. Hansen, and Z. Weng, "Detection of cis-element clusters in higher eukaryotic DNA," *Bioinformatics*, vol. 17, no. 10, pp. 878–889, Oct 2001.

[16] M. C. Frith, J. L. Spouge, U. Hansen, and Z. Weng, "Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3214–3224, Jul 2002.

[17] M. Gupta and J. S. Liu, "De novo cis-regulatory module elicitation for eukaryotic genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 20, p. 70797084, 2005. [Online]. Available: http://www.pnas.org/content/102/20/7079.short

[18] W. Thompson, M. J. Palumbo, W. W. Wasserman, J. S. Liu, and C. E. Lawrence, "Decoding human regulatory circuits," *Genome Research*, vol. 14, no. 10, p. 19671974, 2004. [Online]. Available: http://genome.cshlp.org/content/14/10a/1967.short

[19] Q. Zhou and W. H. Wong, "CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 33, pp. 12 114–12 119, Aug 2004.

[20] D. Papatsenko, A. Kislyuk, M. Levine, and I. Dubchak, "Conservation patterns in different functional sequence categories of divergent drosophila species," *Genomics*, vol. 88, no. 4, pp. 431 – 442, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888754306000796

[21] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct 1990.

[22] A. A. Nikulova, A. V. Favorov, R. A. Sutormin, V. J. Makeev, and A. A. Mironov, "CORECLUST: identification of the conserved CRM grammar together with prediction of gene regulation," *Nucleic Acids Res.*, vol. 40, no. 12, p. e93, Jul 2012.

[23] S. Sinha and X. He, "MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules," *PLoS Comput. Biol.*, vol. 3, no. 11, p. e216, Nov 2007.

[24] S. Sinha, M. D. Schroeder, U. Unnerstall, U. Gaul, and E. D. Siggia, "Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in drosophila," *BMC bioinformatics*, vol. 5, no. 1, p. 129, 2004. [Online]. Available: http://www.biomedcentral.com/1471-2105/5/129

[25] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," *BMC Genomics*, vol. 13 Suppl 4, p. S2, 2012.

[26] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun 2002.

[27] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D52–57, Jan 2011.

[28] E. Portales-Casamar, S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. W. Wasserman, and A. Sandelin, "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles," *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D105–110, Jan 2010.

[29] B. J. Raney, M. S. Cline, K. R. Rosenbloom, T. R. Dreszer, K. Learned, G. P. Barber, L. R. Meyer, C. A. Sloan, V. S. Malladi, K. M. Roskin, B. B. Suh, A. S. Hinrichs, H. Clawson, A. S. Zweig, V. Kirkup, P. A. Fujita, B. Rhead, K. E. Smith, A. Pohl, R. M. Kuhn, D. Karolchik, D. Haussler, and W. J. Kent, "ENCODE whole-genome data in the UCSC genome browser (2011 update)," *Nucleic Acids Research*, vol. 39, pp. D871–D875, 2010. [Online]. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq1017