

QUERYING A DATABASE OF REGULATORY ELEMENTS

CHENGYONG YANG[†] &
ERLIANG ZENG[†]

*BioRG, School of Computer Science
Florida International University
Miami FL 33199, USA
cyang01,ezeng001@cs.fiu.edu*

KALAI MATHEE

*Department of Biological Sciences
Florida International University
Miami FL 33199, USA
matheek@fiu.edu*

GIRI NARASIMHAN

*BioRG, School of Computer Science
Florida International University
Miami FL 33199, USA
giri@cs.fiu.edu*

In this paper, we present a new approach that will facilitate the process of drawing meaningful conclusions that are likely to be useful to a biologist who may be interested in only a specific functional aspect of the organism (rather than the entire genome). This is achieved by constructing a comprehensive relational database and an intelligent, user-friendly query system. With the help of simple examples, we show how a biologist with no prior knowledge of databases can use this system. Our database can be queried from the URL: <http://biorg.cs.fiu.edu/TFBM/>

1. Introduction

Genomic databases have seen a great surge in the volume of sequence data and the number of annotated open reading frames (ORF). Biologists have used experimental approaches to delineate gene regulations for generations. But experimental determination is slow, laborious, and not practical on a genomic scale. To speed up the process of understanding gene regulation, biologists have sought the assistance of bioinformatics approaches to obtain clues that would help them design appropriate experiments. Many software programs have been developed to predict transcription factor binding motifs (TFBM) by applying

[†] Authors Chengyong Yang and Erliang Zeng contributed equally to this work.

sequence pattern discovery tools to nucleotide sequences upstream of the ORFs [7, 9, 12]. Recent methods combine pattern discovery techniques with information from gene expression to predict TFBMs [4].

One of the difficulties with these programs is that they produce a large number of predicted TFBMs along with associated scores representing the statistical significance of the predictions. However, drawing biologically useful inferences or conjectures remains a difficult problem. In this paper, we present a new approach that will facilitate the process of drawing meaningful conclusions that are likely to be useful to a biologist who may be interested in only a specific functional aspect of the organism (rather than the entire genome). This is achieved by constructing a comprehensive relational database and an intelligent, user-friendly query system. With simple examples, we show how a biologist with no prior knowledge of databases can use this system.

In a previous paper, we used upstream sequence data and gene expression data for the organism, *P. falciparum*, and discussed the construction of a database of information obtained from using a variety of programs [15]. *P. falciparum* is one of four species of the parasitic protozoan genus Plasmodium, and is responsible for the vast majority of malaria episodes, affecting 200-300 million individuals and causing 0.7-2.7 million deaths per year worldwide. Elucidating the regulations of genes will be helpful to understand the functions of *P. falciparum* genes during development processes. The goal of that study was to integrate information from various sources, infer relationships between transcription factors (TF) and their targets, and discover novel TFBMs [15]. In that study, we used two different approaches to obtain information about TFBMs and put all the information into a MySQL database. The first approach involved using MotifRegressor to effectively discover expression-mediating TFBMs of medium to long width. MotifRegressor first constructs candidate motifs and then applies regression analysis to select motifs that are strongly correlated with changes in gene expression [4, 9]. The second method involved a combination of the Iterative Signature Algorithm (ISA) [2, 8] to obtain clusters of genes that were potentially co-regulated at specific stages of the development of the parasite, and then applying the AlignACE algorithm to predict potential TFBMs [7]. Then the results were all put into a MySQL database. For the sake of convenience, we will refer to this database as PlasmotFBM.

The resulting database contained:

1. All the discovered TFBMs, along with their significance scores, the software using which they were found, and the genes whose upstream

sequences contained them along with their location in those upstream sequences.

2. Clusters of co-regulated genes (referred to as *transcription modules*, or simply *modules*), and the time points at which they were co-regulated.
3. All genes and ORFs in the genome, their chromosomal location, their functional annotation, and their expression information at all the time points during the development of the parasite.

As a proof-of-concept we showed that many known *P. falciparum* motifs were identified by our methods.

In this paper, we started with the above database, PlasmoTFBM, and focused on building a user-friendly, web-based query and visualization system to help a user to ask non-trivial questions. The user may be a biologist, who may wish to investigate some biologically meaningful question and may have no knowledge of designing queries to a sophisticated database system. We will refer to our query system as QTFBM.

In Section 2, we briefly describe some of methods used in this paper. In Section 3, using simple examples, we demonstrate the capabilities of the QTFBM system. We conclude with some discussions in Section 4.

2. Methods

Figure 1 shows a schematic of the procedures that were used to build the PlasmoTFBM database. As shown, information from the gene expression data (QC data) and the upstream sequences of the genes was used to generate a variety of modules and motifs using several different software packages (ISA, AlignACE, and MotifRegressor). All this information was then entered into the PlasmoTFBM database.

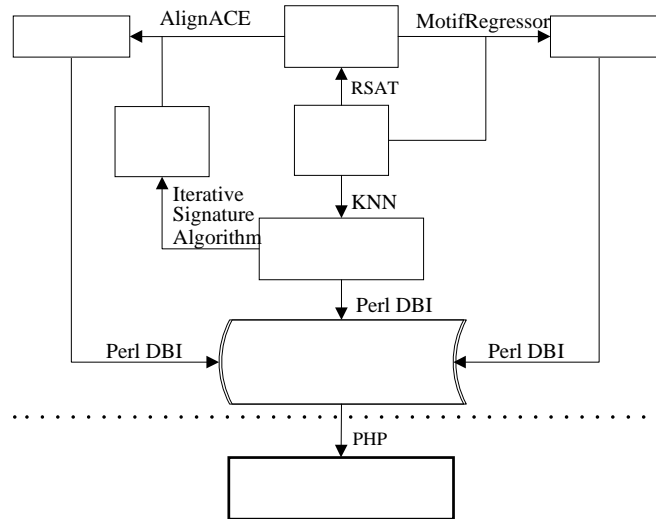


Figure 1: Flowchart for mining TFBS for *P. falciparum*.

2.1. Generating potential TFBS

Quality gene expression data (QC data) was downloaded from the Malaria Transcriptome Database [<http://malaria.ucsf.edu/SupplementalData.php>].

Missing values were estimated using the K-nearest neighbors (KNN) method implemented in R [13]. The corresponding upstream sequences were extracted by regulatory sequence analysis tools (RSAT) [14]. The ISA algorithm was applied on available collections of related genes. In total, 217 modules were obtained, with gene sets ranging in size from 10 to 500. The resulting modules were assumed to be co-regulated gene clusters, and used as input to the AlignACE software to generate a set of motifs [<http://atlas.med.harvard.edu>] common to these clusters. In parallel, the MotifRegressor software was run on the gene expression data to obtain significant motifs. Motifs that were identical, similar, or overlapping with the motifs discovered by AlignACE were merged using Perl scripts (the cleaning step). This resulted in 424 motifs generated from MotifRegressor and 1249 from AlignACE.

2.2. Database

A relational database, referred to as PlasmoTFBS, was designed and implemented using MySQL to store all the available information. As explained before, this database includes the gene expression data, generated significant

motifs and modules, gene annotation information including the functional information and the chromosomal location.

2.3. Web Query and Visualization

Web query interface was implemented using PHP. As shown previously [15], it is possible to design complex queries for the PlasmoTFBM database using PerlDBI. In fact, we used such a method to successfully analyze motifs in the gene EBA140 [15]. While such methods can be used to analyze motifs, it requires non-trivial expertise to be able to use it effectively. This motivated us to build a handy web-based query system that could automate some of the analyses.

The motivation for the query system was as follows. Most biologists perform research on a small set of genes, usually a set of genes that are involved in a specific function or a specific pathway. Such a biologist would be interested in knowing whether this database has results that are relevant to their genes of interest, i.e., what other genes are co-regulated with the ones in questions, what motifs they might share, what developmental stage or functional pathway they might be involved in, what transcriptional factors may be regulating the genes of interest, and finally, what biologically meaningful conjectures can result from the analyses and that may be relevant to the genes of interest. Answers to such questions may be the starting point for further investigations for the biologist.

Consider the following example. Assume that the genes of interest are MAL13P1.60 and MAL7P1.86. MAL13p1.60 encodes the protein erythrocyte-binding antigen 140 (EBA140), which is implicated in merozoite invasion using a sialic acid-dependent receptor on human erythrocytes [1]. MAL7P1.86 is a putative alpha subunit of transcription initiation factor IIE. Furthermore, both genes are expressed highly during the merozoite stage of the parasite's development [3]. A biologist may be interested in studying their relationship: Are they co-regulated (i.e., is there a module that contains both of them, is there a set of time points or conditions under which their expression profiles are correlated)? Do they share any motifs in their upstream regions? Can any other relationships be conjectured?

In the first step, all modules that include some subset of the genes of interest are computed sorted by the number of genes of interest that they contain. Next, the user may choose a subset of the generated modules for further exploration. Suppose that the user decides to explore the module MAL7P1.86_g2_c6, which contains both the genes of interest. The query system also outputs the conditions and gene sets associated with the selected module. The user could then ask for

the list of all the motifs found in the selected modules and under the selected conditions.

Visualization tools were used to visualize the final results in a more meaningful way. All motifs of interest are displayed using the Logo notation. The user may select a specific set of genes and the motifs for each gene of interest is then displayed in a graphical manner by showing their location as a function of their distance from the translation start site (ATG) in the gene upstream sequence. See Figure 4 for an example. The average gene expression profile along with the standard deviation is also displayed for the selected genes. All images are generated dynamically using PHP and the GD library. Thus the web interface makes it possible to mine information and visualize some interesting results simply through a series of mouse clicks, and without the user having to learn a complicated database or a query language.

3. Experimental Results

A small number of regulatory elements in *P. falciparum* have been biologically verified and have been reported in the literature [6]. We sought to validate our results using these known motifs. Below, we also discuss some of the interesting motifs found in our database.

3.1. Investigating the SERA gene family

The SERA family owes its name to the serine repeat antigen, a protein possessing a characteristic serine homopolymer [10]. In *P. falciparum*, they are expressed when the parasite is developing within the erythrocytes. Eight members of the SERA family showed enhanced expression during the mid-schizont stage of the parasite. They are: PFB0325c PFB0330c PFB0335c PFB0340c PFB0345c PFB0350c PFB0355c PFB0360c. These eight genes, all located on chromosome 2, are arranged in tandem. These eight genes are highly co-expressed, and are potentially co-regulated, suggesting that they may share common regulatory elements. Starting with these eight genes, we applied the procedure described in the previous section. A total of eight modules containing a large subset of the eight genes were identified (See Table 1). Among them, two modules (PFE0415w_g1.3_c8 and rand5-80_m22_1) contained five of the eight genes (See Table 2).

Table 1: Modules containing some of the SERA genes

Genes of Interest	Module	Module Condition
PFB0325c	rand5-80_m22_1	1-6, 8-22, 24-28, 30-48
PFB0330c	PFE0415w_g1.3_c8	16, 30-33, 35-37
PFB0335c	rand5-80_m21_1	1-6, 8-22, 24-28, 30-48
PFB0340c	rand5-80_m20_2	1-5, 8-22, 24-28, 30-48
PFB0345c	m12_1	1-5, 7-22, 24-28, 30-48
PFB0350c	rand5-80_m19_1	1-5, 7-22, 24-28, 30-48
PFB0355c	m11_1	1-5, 7-22, 24-28, 30-48
PFB0360c	rand5-80_m18_1	1-5, 7-22, 24-28, 30-48

Table 2: Genes contained in modules PFE0415w_g1.3_c8 and rand5-80_m22_1 and the corresponding motifs significant for the conditions relevant to the module.

Module	Gene	Motif
PFE0415w_g1.3_c8	PFB0330c	Motif.P35.6.3BG, Motif.P31.5.22BG, Motif.P33.5.10BG
	PFB0335c	Motif.P35.6.3BG
	PFB0345c	Motif.P29.5.3BG
	PFB0350c	None
	PFB0360c	Motif.P35.6.3BG, Motif.N29.6.15BG, Motif.P31.5.22BG
rand5-80_m22_1	PFB0340c	Motif.P35.6.3BG, Motif.P42.5.22BG, Motif.P42.5.18BG, Motif.N37.11.15BG, Motif.P37.5.2BG, Motif.N35.11.22BG, Motif.N29.11.25BG, Motif.N30.8.26BG, Motif.P40.6.20BG, Motif.N40.5.11BG, Motif.P42.13.13BG
	PFB0340c	Motif.P35.6.3BG, Motif.N41.11.19BG, Motif.P42.13.12BG, Motif.N40.11.17BG, Motif.P40.6.20BG
	PFB0345c	Motif.P38.7.25BG, Motif.P30.5.4BG
	PFB0350c	Motif.N40.7.27BG

Further investigations showed that the motif Motif.P35.6.3BG was found to be significant in both the modules and to be present in the genes PFB0330c, PFB0335c, PFB0340c, and PFB0360c. Thus we may conjecture that Motif.P35.6.3BG is an important regulatory element in the SERA family. Our query system also showed that Motif.P35.6.3BG also exists in a group of other genes in module PFE0415w_g1.3_c8, and is not just limited to the SERA gene family. This gene group shares the common regulatory element shown below

using the Logo format (Figure 2), and a similar expression profile (See Figure 3), suggesting that they may be co-regulated by a common transcription factor.



Figure 2: Motif.P35.6.3BG found in several SERA genes

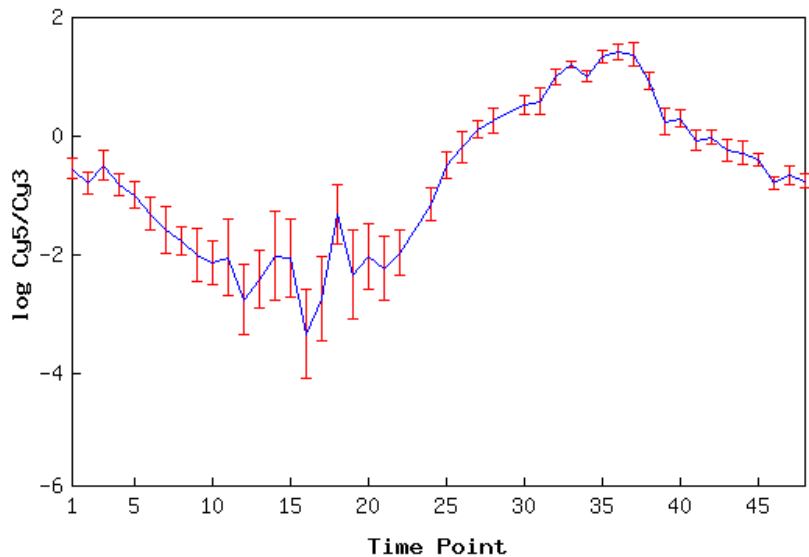


Figure 3: Average expression profile (shown in blue) for module PFE0415w_g1.3_c8. The red bars indicate standard deviation.

It is interesting to note that module PFE0415w_g1.3_c8 contains a putative transcription factor PFE0415w, suggesting that some genes in this module may be regulated by it. Figure 3 shows the average expression profile for this module, which can be visualized using our query system. The average profile shown in Figure 3 may be used to see that the standard deviation for the profile is small during the time period 25 through 48, showing that the genes must be tightly co-regulated during this period.

Our web-based query system also shows the motif locations on the genes of interest in a graphical manner. See Figure 4 for an example.

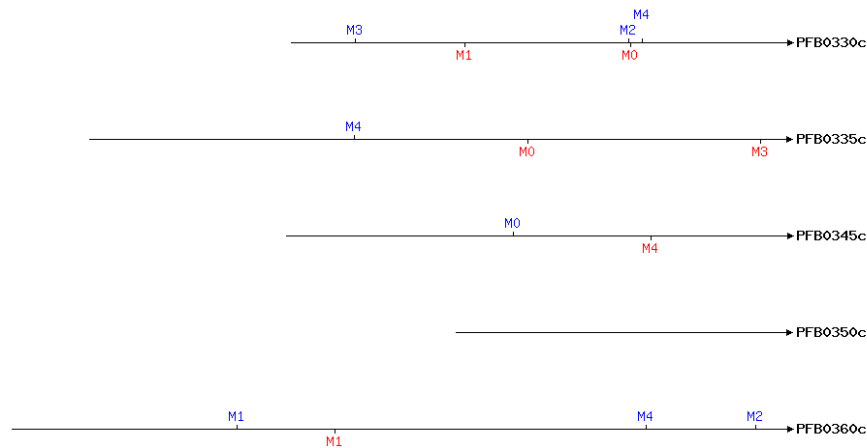
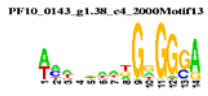
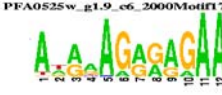

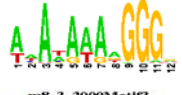
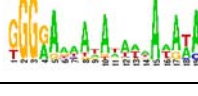


Figure 4: Motifs visualized in the upstream sequence of all the SERA genes of interest. The line indicates the upstream sequence with the translation start site at the right end. Motifs labeled as M0, M1, M2, M3, and M4 correspond to motifs Motif.P29.5.3BG, Motif.N29.6.15BG, Motif.P31.5.22BG, Motif.P33.5.10BG, and Motif.P35.6.3BG, respectively. Red color represents motifs in the forward direction while blue color represents those in the reverse direction.

3.2. Known Motifs

As mentioned before, some transcription factor binding sites have been previously reported in the literature. One particular TFBM of interest is the G-box present in heat shock proteins (hsp) [11]. The G-box regulatory element is considered to be unique to *P. falciparum*, since no homologous eukaryotic element is known. Militello *et al.* only report the G-box from hsp genes. However, querying our database shows that the G-box motif and its variants exist throughout the whole genome and may be present in genes involved in a wide variety of functions. The matches to motifs in our database are listed below in Table 3. Detailed results will be available as supplemental materials on our website [<http://biorg.cs.fiu.edu/TFBM/>].

Table 3: G-box motifs appearing in the upstream sequences of the hsp genes are given in the first column. The motifs shown using the Logo format were obtained by using AlignACE on the modules listed in the second column.

Locus	Module	Logo [5]
PFI0875w (HSP)	PF10_0143_g1.38_c4	
PF11_0175 (HSP 101)	PFA0525w_g1.9_c6	
PF11_0188 (HSP 90)	m8_5	
PFL0740c (hypothetical)	m11_2	
PF08_0032 (hypothetical)	m8_3	

Other motifs reported in the literature that were found in our database include the cis-acting element in CDP-diacylglycerol synthase gene promoter, the CPE motif in *var* genes 5B1 upsC, and the SPE2 motif in all upsB *var* gene sequences. Refer to supplemental materials in our website for details.

4. Discussion and Conclusions

In the previous section, we used the SERA gene family as an example to elucidate how to use our web-based query system to mine for biologically meaningful information. The example shows that our interface is easy to use for a biologist and has the ability to arrive at useful conjectures. Any gene set of interest can be provided as input in order to do the analysis.

We used the *P. falciparum* as a model system on which to try our experiments. Since *P. falciparum* has two life cycles: sexual and asexual, and two hosts: human beings and the female Anopheline mosquito, its regulatory machinery is believed to be considerably different from that of other eukaryotic organisms. The database we built can be used to compare *P. falciparum*

regulatory machinery with that of other eukaryotic organisms and can be combined with other databases containing eukaryotic regulatory elements.

Another application of this database is to classify the genes involved in specific biological process based on their expression profile and the regulatory element present in their upstream regions. For example, DeRisi's lab reported 58 ORFs to be important for the invasion process during the merozoite stage [3]. Querying our database using these 58 ORFs produced many modules under many different sets of conditions. Further investigations are needed to study the results from these analyses, and to determine significant motifs shared by them.

For *P. falciparum*, only 14 transcription factors have been identified so far. In future, if more transcription factors are determined, then this database could be used to obtain putative binding sites and genes potentially regulated by it. This can help the biologist to design further experiments to verify and validate the predictions and conjectures.

Acknowledgements

Authors CY and EZ contributed equally to this paper. We thank Haifeng Wang for helpful discussions. We also thank Jing Zhai and Wei Shi for help with Logo. EZ is supported by a Florida International University Presidential Graduate Fellowship. Research of GN was supported in part by NIH Grant P01 DA15027-01.

References

1. J. Baum, A.W. Thomas and D.J. Conway, *Genetics*, **163** (4), 1327 (2003).
2. S. Bergmann, J. Ihmels, and N. Barkai, *Phys. Rev. E*, **67**, 031902 (2003).
3. Z. Bozdech, M. Llinas, B.L. Pulliam, E.D. Wong, J.C. Zhu, and J.L. DeRisi, *Plos Biology*, **1** (1), 85 (2003).
4. E.M. Conlon, X.S. Liu, J.D. Lieb, and J.S. Liu, *Proc Natl Acad Sci U S A*, **100** (6), 3339 (2003).
5. G.E. Crooks, G. Hon, J.M. handonia, and S.E. Brenner, *Genome Res*, **14** (6), 1188 (2004).
6. P. Horrocks, K. Dechering, and M. Lanzer, *Mol Biochem Parasitol*, **95** (2), 171 (1998).
7. J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church, *J Mol Biol*, **296** (5), 1205 (2000).
8. J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, *Nat Genet*, **31** (4). 370 (2002).
9. X.S. Liu, D.L. Brutlag, and J.S. Liu, *Nat Biotechnol*, **20** (8). 835 (2002).

10. O. Mercereau-Puijalon, J.C. Barale, and E. Bischoff, *Int J Parasitol*, **32** (11), 1323 (2002).
11. K.T. Militello, M. Dodge, L. Bethke, and D.F. Wirth, *Mol Biochem Parasitol*, **134** (1), 75 (2004).
12. F.P. Roth, J.D. Hughes, P.W. Estep, and G.M. Church, *Nat Biotechnol*, **16** (10), 939 (1998).
13. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman, *Bioinformatics*, **17** (6), 520 (2001).
14. J. van Helden, *Nucl. Acids. Res.*, **31** (13), 3593 (2003).
15. C. Yang, E. Zeng, K. Mathee, and G. Narasimhan, *Proc. of CAMDA'04*, (2004).