Searching Fly mRNA Sequences for Motifs for Enhanced piRNA Levels at the 3′ UTR of Those Sequences

Eugene Wolfson's COSI 178A Term Project

May 9, 2011

Abstract

A library of *D. melanogaster* mRNA sequences is split into a positive and negative set by the level of piRNA mapped to their 3′ UTR. The resulting positive and negative set is used to run various *de novo* motif discovery algorithms. The discovered motifs are used as features in machine learning algorithms in order to evaluate their statistical significance and discover possible combinations of candidate motifs that lead to enhanced piRNA levels at the 3′ UTR. The strengths and weaknesses of this approach are weighed, and the results from several motif discovery algorithms are reported and analyzed.

Introduction

Piwi proteins preferentially bind to a class of short RNAs called Piwi-interacting RNAs (piRNAs). piRNA is between 24 and 31 nucleotides long. In previous work, Lau et. al. mapped highly enriched levels of piRNA back to the $3'$ UTR of a set of genes we will refer to as the positive set. In this work we try to find a motif or set of motifs to explain the enriched piRNA levels.

Before searching for motifs, we note the directionality of the piRNA, because using this information, we can gain some insight into the process by which it was generated. Since proteins binding to DNA express both of its strands, we would expect piRNA to be found equally in either direction if it came from DNA; however, the piRNA that we found has a singular directionality. This observation gives rise to the hypothesis that the piRNA we observed could be generated from mRNA, which only has one strand, and thus can be expressed in only one direction.[1]

In order to test the hypothesis that the piRNA comes from mRNA, we try to find motifs within the mRNA sequences from the positive set, which are not enriched in the negative set. Finding motifs to account for the positive set could give some backing to the hypothesis and provide further insight into the nature of piRNA. A library of mRNA sequences from *D. melanogaster* germline cells was sequenced. The level of piRNA mapped back to the $3'$ UTR of each mRNA sequence was noted. The sequences were then divided into a positive and negative set using a threshold, with genes in the positive set having a much higher level of piRNA than those in the negative set, which have a negligible level of piRNA.

There were two data sets created from this library. The first data set has 1315 sequences in the positive set and 276 sequences in the negative set. No piRNA information besides its binary enrichment status is available in the first data set. The second data set has 223 sequence in the positive set and 250 in the negative set. The piRNA level for each sequence is available: the positive set spans three orders of magnitude of piRNA levels, from as low as 73 to over 73 thousand; the negative set's piRNA levels are very low, ranging between 2 and 9. Unless mentioned otherwise, our analysis is on the first data set.

A brief overview of *de novo* motif discovery follows. We can use computational techniques to find candidate motifs. Those candidate motifs found to be significant can be biologically verified; for example, some part of the motif can be knocked out on a positive gene and the resulting piRNA level can be compared to that of the wild-type.

In the simplest case, the presence of a sequence motif is enough to classify a gene as positive. More likely, the piRNA level depends on a more complex scenario. For example motifs A and B must be present, or maybe A and B need to present, but C should not be; or A must be a certain number of base pairs downstream of B. The point is that the search space is very large and complex

It could even be the case that there is no set of sequence motifs associated with piRNA level. We just don't know enough about the biology to say. Nonetheless, finding a motif would help us understand the process by which piRNA is created. It should be noted that in general, a motif can be any feature or pattern that is conjectured to have a biological significance, but in this paper unless specified otherwise, a motif usually refers to a short genomic sequence that can be expressed as a position-specific scoring matrix (PSSM). A PSSM is a $W \times A$ matrix, where W is the length, in nucleotides, of the motif and A is 4, the length of the genomic alphabet. Each cell of the PSSM matrix is $\log_2[\theta_{i,j}/\theta_{0,i}]$, where $\theta_{i,j}$ is the probability of the letter i occuring at position j of the motif and $\theta_{0,i}$ is the background genome background probability of the letter i. We devote the rest of this paper to discussing the approaches to and preliminary results of the search for candidate motifs.

Methods

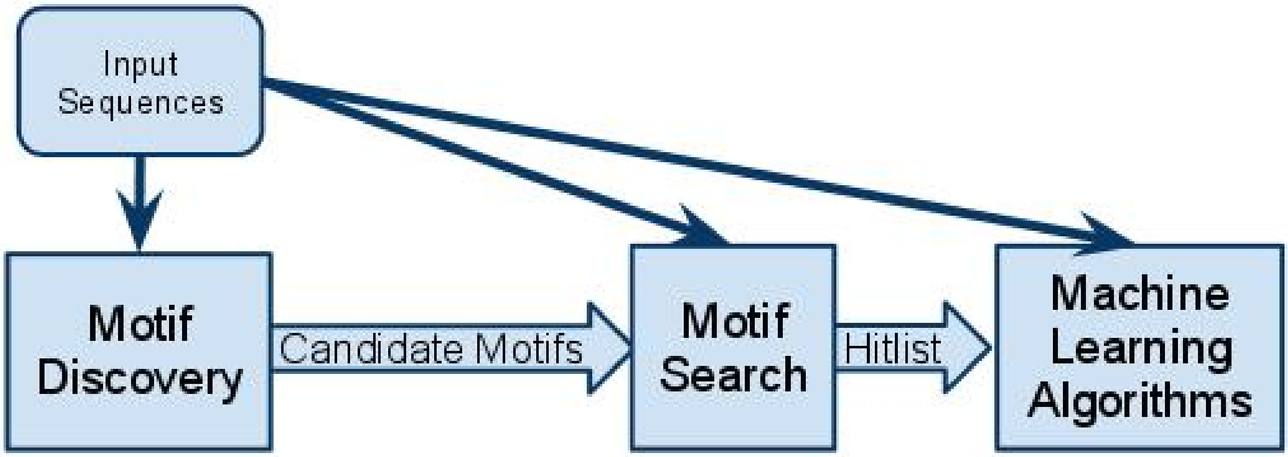

Figure 1: Motif Analysis Pipeline. Each step is modular and takes a list of sequences as input (in different formats). The order of the steps is (1) Motif Discovery, (2) Motif Search, and (3) Analysis via Machine Learning.

No one motif discovery tool can accurately predict motifs, so it is often helpful to use several tools and analyze the results in some measurable way.[2] In this section, we present the methods used to analyze the candidate motifs we find using various motif discovery tools. Figure 1 is a visualization of the three-stepped motif analysis pipeline we describe.

Motif Discovery

The first step of pipeline is motif discovery. A variety of motif discovery algorithms such as MEME, AlignACE, Weeder, et cetera are used. [3, 10, 9]

The common input to each is a list of sequences from the positive set. Some tools also accept a negative set as an optional parameter. In the conclusion section, we mention a potential issue arising from using the negative set, but using the negative set can help filter out false positives. The sequences are mRNA sequences, which are single-stranded; therefore, no reverse complement (also know as negative) strand exists and the appropriate flags should be set on each tool.

The output is a set of candidate motifs in various formats depending on the program used. Each candidate motif should be converted to a PSSM for the Motif Search step.

Motif Search

After a set of candidate motifs is identified, we want to identify where in our positive and negative set the motifs are found. In our analysis, MAST[4] was used, but any simple motif detection tool works. For MAST, the input is a set of MEME formatted PSSMs of the candidate motifs, as well as the positive and negative set of sequences to be searched. MAST outputs a list of where each motif is found in each sequence, along with a score and p-value. Those matches that are considered statistically insignificant, or false positives, should be ignored. The rest should be converted into a hitlist to be used in the Machine Learning Analysis step. MAST has flags for automatically converting only the best matches to a hitlist.

Machine Learning Analysis

The last step in the pipeline is the analysis of our candidate motifs using the hitlist. We used the machine learning framework called WEKA[5] for our analysis. WEKA takes a file in .arff format as input. This file is a labeled feature list for each sequence. That is, for each sequence there is information about the value of each feature for that sequence, and each sequence has a label.

Features can be anything describing the sequence in question. An example of a feature is the presence of a specific candidate motif in a sequence. The value of each feature can be either binary or numeric. A binary feature could be whether a certain motif was found in a sequence or not; on the other hand, a numeric value could give the score that the motif was assigned by MAST. The features do not have to be limited to sequence motifs. Other examples include the presence of a certain structural motif, or the number of sequential repeats a specific codon (Glutamine GA[AC]) has in a sequence.

Each label should be either the name of the set that the motif belongs to, i.e. "positive" or "negative"; or a numeric score such as the piRNA level mapped to the $3'$ UTR of the sequence.

The machine learning algorithms use this .arff formatted file to create a model, which we then analyze to make decisions about our set of candidate motifs. Basically, the goal of this step is to find a set of features that helps classification, along with information about any inter-feature interactions. The output will depend on the machine algorithm, or classifier, we use and upon the quality of the input features.

Before we make decisions about the candidate motifs, we should maximize the accuracy of the classifier. This can be done by adjusting parameters, finding helpful features, and providing more and better data in the form of sequences. Only once we are satisfied with the quality of our classifier, should we analyze the model it creates for insights about our candidate motifs.

It should be noted that training the classifier is very quick, because input set size is relatively small compared to the tasks that ML is usually used for: 10^3 sequences, each with 10^1 features (motif information) makes up an input set of 10^4 nodes. As an example, creating a regression tree (WEKA's REPTTree) takes less than a minute on a modern computer with a 1.8 GHz processor.

Results

Candidate Motifs Discovered

When we ran MEME on 50 random sequences from the positive set, constrained to motifs between 7 and 11 base pairs, we discovered a single length 11 motif that consists of a repeating Glutamine codon. The amino acid Glutamine is coded by either CAG or CAA. Searching the top 50 sequences from the positive set, with length 6-10, the top motif was CAG repeating. This motif is similar to, but not the same as the length 11 repeating Glutamine motif. CAG still codes for Glutamine, but CAG is different from CA[GA]. Searching the top 50 sequences from the positive set, with length 6-8, we found the six motifs in figure 2c. Note that these were created using both strands, which is biologically incorrect; however, the motifs coming from reverse complement strands, such as the motif with the regular expression TTTTTTGT[TG], will be found to be insignificant in the Motif Search and Machine Learning steps, so we can still consider these results in the set of candidate motifs. Note that the repeating Glutamine codon comes up in figure 2c as two separate motifs: the top left motif is repeating CAA and the bottom center motif is repeating CAG.

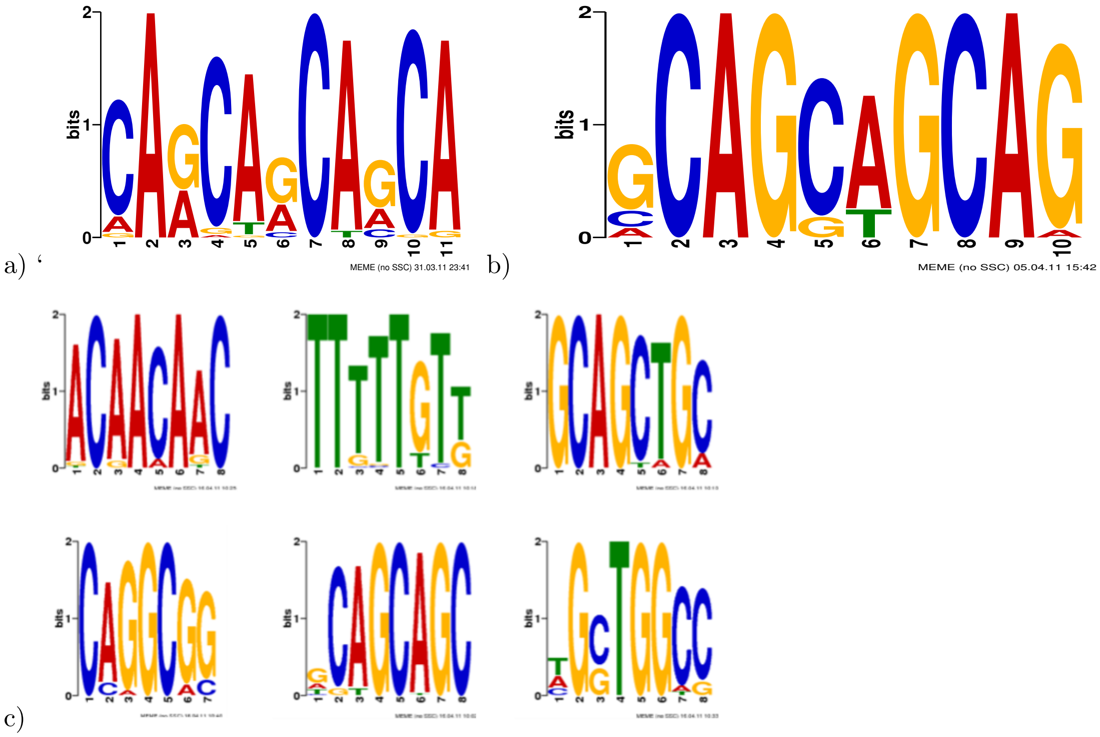

Figure 2: a) Repeating Glutamine Motif of length 11. Created using a random 50 sequences from the positive set. b) Repeating GAC of length 10. Created with the following command: `./meme top50.fasta -minw 6 -maxw 10 -allw -dna` c) The top six motifs of length 7 and 8. Created with the following command: `./meme top50.fasta -minw 6 -maxw 8 -allw -dna`

Despite its appearance in MEME's results, Glutaime is a common amino acid found in many genes that do not express piRNA. Due to this biological intuition, we are not convinced that a repeating Glutamine codon is a motif for increased piRNA levels in a gene. We tried to verify the repeating Glutamine motif by seeing if it could be detected by algorithms other than MEME. The two other programs we tried, AlignACE and Weeder, did not find a repeating Glutamine motif, but did turn up other motif candidates. Both AlignACE and Weeder were run using the top 50 motifs from the positive set. AlignACE turned up one hundred motifs, most of them long and insignificant. We discounted all the motifs that had a length greater than 11. The logos of the five remaining motifs are in Table 1. Weeder returned the 11 motifs in Table 2, but none of them turned out to be significant when the machine learning analysis was applied to them.

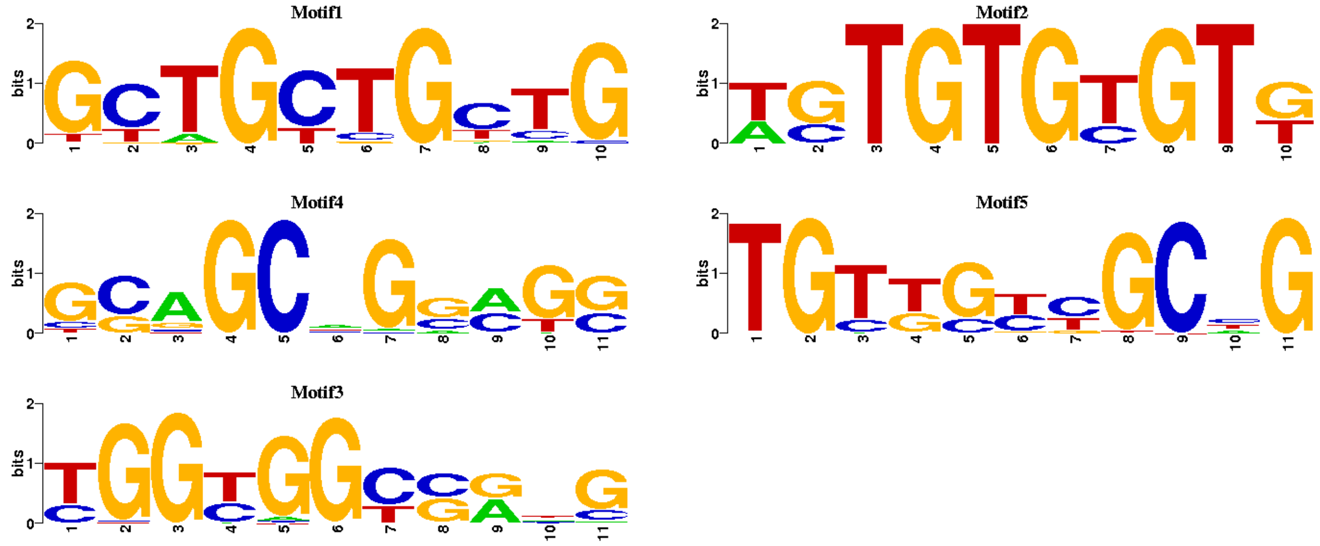

Table 1: AlignACE motifs

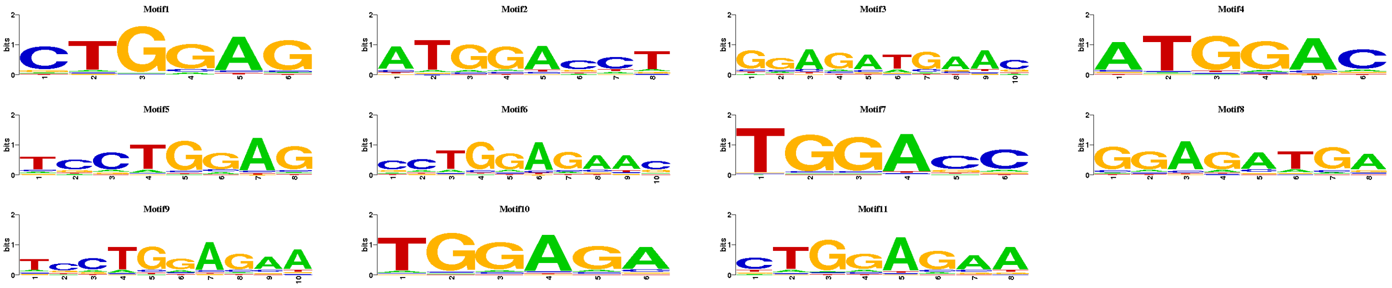

Table 2: Weeder motifs

WEKA analysis

The analysis of the first data set suffered from data imbalance. The positive set was much larger than the negative set, so for some sets of candidate motifs, the resulting decision tree often defaulted to classifying everything as positive. This most likely means that the features under consideration were not helpful. The reason for defaulting a positive classfication is that machine learning algorithms try to maximize their correctness, and classifying everything as belonging to the largest set is a naive but effective strategy. In the case of two classes, it will be right most of the time, by definition; and if the given features are not helpful enough, it could be the best possible strategy. This issue can be avoided by weighing the negative set proportionally to the positive set, or making sure that both sets are of similar size.

We ran the candidate motifs we found using the first data set through the Motif Search and Machine Learning steps of the pipeline, but with input sequences from the second data set. Overall, the accuracy of the ML step seemed to be worse when using the second data set than it was with the first. Having a smaller overall data set may have been responsible for decreasing the accuracy of the results. A portion of the sequences in the second data set were different from the first data set. Other than the data imbalance issue of the first data set, one possible reason that the first data set performed better is because of overfitting. That is, the candidate motifs were discovered using the positive subset of the first data set, and in effect are guaranteed to appear on positive subset of the first data set. There is no guarantee that these candidate motifs will appear on the sequences in the second data set that are not present in both data sets.

The second data set contains an almost equal amount of sequences with the positive label as it has sequences with the negative label. The second data set also has numeric piRNA level information, which is more precise than the binary class label ("positive" or "negative"). Using the score of the hits on the hitlist instead of using the binary "hit"/"no hit" seemed to slightly increase the accuracy of the classifier. Judging whether using the numeric piRNA instead of a binary class label, is harder. Creating a confusion matrix is not possible if the label is numeric. It can be said that the decision tree of numerically labeled sequences was much larger, implying that more complex patterns can be discovered; however, I did not know how to statistically compare the results without a confusion matrix or an F-score - so the impact of using a numerical label instead of a binary one remains unstudied.

Conclusion

Discovered Motifs

The results reported in this paper are very preliminary but we do have some suggestions for the work going forward.

We recommend searching for motifs of length 8. Length 8 seems to be long enough to encapsulate the information about the potential motifs - if any exist, because no significant motifs were found with a length higher than 8, that were also not found with a length of 8; finally, eight is short enough to keep the search from being impractically slow, even if the whole data set is used. As a side note, some motif discovery tools only accept even motif lengths.

Only three motif discovery programs were tested. A greater variety of algorithms should be tested. A program for running motif discovery algorithms in bulk, called Tmod[6] exists. Tmod provide an interface to most commonly used motif discovery algorithms and can reduce the hassle of installing each program.

Only a subset of the positive set was used for motif discovery. The reason for this was to limit the run time. There is a discussion of practical run times for MEME at the NBCR forum. Basically, “the MEME algorithm run time is cubic with respect to the number of input sequences,” and “quadratic with respect to the number of characters.” Despite the increased CPU time, using the whole data set might have helped discover more enriched motifs.

The motif discovery algorithms used in the first step should be tested systematically. This can be done by inserting an aritificial sequence motif into the positive sequences set, to see which discovery algorithms could detect it. This test would also work for evaluating the effectiveness of the machine learning step and of the motif search step.

Although the repeating Glutamine motif was not confirmed by AlignACE or Weeder, our investigation by no means has discounted the possibility that a repeating Glutamine codon increases piRNA levels in the fly gonad. Several more tests should be run. Most obviously, other discovery algorithms should be run on the data. Also, a hypothesis has been presented by Gung-wei Chirn: the number of repetitions of the subsequence could play a role in determining piRNA level. This experiment hasn’t been run, but it could easily be tested by plugging the number of GA[CA] repetitions as a feature in the input to the machine learning step of the pipeline.

In addition to sequence motifs, RNA has secondary structure. We tried to find structural motifs in the top 50 sequences from the positive set using rnamotifs08[7], but could not find any results. Finding structural motifs for RNA is still a very new field, but it’s worth a second glance.

Pipeline

The pipeline described in this paper is designed to detect false positives identified by *de novo* motif discovery algorithms - candidate motifs that are not significant, but are incorrectly put in the result set. However, the problem of false negatives persists. False negatives are subsequences that should be in the set of candidate motifs but are missed during motif discovery. Here we address some sources of false negatives and possible solutions.

Since the motif discovery algorithms used try to find the most enriched motifs, it is likely that we miss some candidate motifs that were enriched only in a relatively small subset of the data. A way to work around this issue is to use the results of the machine learning step to facilitate a bootstrapping approach to motif discovery. Basically, we will run the pipeline described in the method section multiple times, removing a subset of the positive sequence on each run. Specifically, on each run, identify the sequences that appear on the hitlist of the most common motifs. In the beginning of the next run, remove these matched sequences from the consideration of the motif discovery step. Continue until the input sequences list is empty or no candidate motifs are found.

An obvious problem is that machine learning approaches work better with more data, but we are reducing the number of sequences in our training set with each iteration. If the motif discovery algorithm allows it, we should opt for providing it with a blacklist of motifs at each iteration, instead of limiting the sequence data.

Since the motif discovery algorithms we used try to find individual motifs, it is likely that we missed some candidate motifs that only work in conjunction with one another. For example, if subsequences A and B appear in both the positive and negative set, but only appear together in the positive set, then the presence of both A and B could be a motif for increasing piRNA levels. However, we may run into problems identifying A and/or B as candidate sequence motifs. Most of the motif discovery algorithms we mentioned take the negative set as an optional parameter to filter out subsequences that appear in both sets. Omitting the negative input sequences from the motif discovery step would give us more false positives, but would potentially reduce the number of false negatives occuring due to the problem described above. Remember that a false negative in the motif discovery step means that a true candidate motif is not added to the set of candidate motifs.

One last point to consider is the choice of using MAST in the motif search step. MAST is a rather complex algorithm, while motif search is a rather straightforward step. It's possible that using FIMO[8] would be more appropriate. According to the NBCR forum, "FIMO is looking for the best individual matches to motifs. MAST is looking for sequences the have the best overall match to a collection of motifs. FIMO's task is simple: given a set of motifs and a database of sequences, compute the match score to each motif at each position in each sequence, and report all the motif matches that pass the p-value/q-value threshold. MAST's algorithm is more complex. For each sequence it carries out an initial scoring that is quite similar to FIMO's. MAST then picks the best match for each motif in the sequence. The p-values of these top matches are multiplied together to create an overall score for the full sequence. MAST reports the sequences that have the most significant overall scores. Typically MAST would be used to look for regulatory regions in DNA, or structures in proteins, where several motifs might occur near each other." MAST may do too much work and omit hits that would be important for the Machine Learning to know about. We could use MAST as further verification of a set of significant candidate motifs that have passed the third step, but FIMO seems to better better fit for the Motif Search task since it considers each candidate motif independently.

The author would like to thank Gung-wei Chirn, Pengyu Hong, and Nelson Lau for their help, explanations, and advice in all aspects of this project.

References

[1] DOI 10.1016/j.cub.2009.11.064 - A Broadly Conserved Pathway Generates 3' UTR-Directed Primary piRNAs

[2] doi:10.1038/nbt1053 - Assessing computational tools for the discovery of transcription factor binding sites

[3] Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

[4] Timothy L. Bailey and Michael Gribskov, "Combining evidence using p-values: application to sequence homology searches", Bioinformatics, Vol. 14, pp. 48-54, 1998

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[6] Hanchang Sun, Yuan Yuan, Yibo Wu, Hui Liu, Jun S. Liu and Hongwei Xie. Tmod: Toolbox of Motif Discovery. Bioinformatics 2009; doi: 10.1093/bioinformatics/btp681

[7] http://genie.weizmann.ac.il/pubs/rnamotifs08/rnamotifs08_exe.html

[8] FIMO. http://meme.nbcr.net/meme4_6_1/fimo-intro.html

[9] Weeder. http://159.149.109.9/modtools/

[10] Finding DNA Regulatory Motifs within Unaligned Non-Coding Sequences Clustered by Whole-Genome mRNA Quantitation, Roth, FR, Hughes, JD, Estep, PE & GM Church, Nature Biotechnology 1998 Oct;16(10):939-45.