GASP: Genetic Analysis of *S. Pombe*

Jake Martinez

1 Introduction

Many current motif finding pipelines attempt to discover motifs through the use of multiple complementary tools. Comparison and contrast in the results aids in determining additional information about the motifs that are being searched for. Our project is primarily concerned with the discovery of regulatory motifs in the promoter regions of six specific genes in fission yeast *Schizosaccharomyces pombe*. These six genes – Mad1, Mad2, Mad3, Bub1, Bub3, and Mph1 – are commonly associated with the spindle assembly checkpoint (SAC). Furthermore, similar genes are found in four other related yeast species: *Saccharomyces cerevisiae*, *Schizosaccharomyces cryophilus*, *Schizosaccharomyces octosporus*, and *Schizosaccharomyces japonicus*; as well as in two other species: *Homo sapiens*, and *Drosophila melanogaster*.

The SAC is vital to the signaling pathway due to its protection of the integrity of the genome through detection and response to errors in chromosome attachment prior to the anaphase [1]. Any defects in spindle structure or alignment of chromosomes on the spindle can result in the death of cells, in the best case, to catastrophic effects such as birth defects in humans or cancerous cell growth [1]. Our goal for this project is to determine the regulatory motifs of each of these genes in order to better understand and potentially control the workings of the SAC. This can be accomplished through interspecies genetic analysis of each of the genes individually as well as a more general analysis of each of the species individually.

2 Methods

The GASP project primarily relies on the benefits of our pipeline. The fundamental aim of the pipeline is to utilize existing tools to accelerate the development cycle of the motif finder in order to focus more heavily on analysis and testing different data sets. The analysis and testing of the results produced by the pipeline are then used to reevaluate and assess the success of scoring mechanisms within the pipeline itself, in an attempt to improve future results.

2.1 Pipeline

For our pipeline, we decided to use five different motif finding tools. Each of these tools have a specific unique component which we decided would assist in the comprehensively of our pipeline. Directly following the completion of each of the five different tools, the results of each tool are coalesced for further analysis and subsequent visualization [1].

[1] refer to Appendix 6.1 for the pipeline diagram

2.1.1 Motif Finding Tools

MEME. The MEME tool is one of the most widely used motif discovery tools in use. It has a vast suite of additional tools that complement the MEME tool itself, providing a great deal of functionality and practicality. One of the main distinguishing factors of MEME from many of the other tools we included in our pipeline is the unique algorithmic approach; MEME utilizes a specialized form of the expectation maximization algorithm to elicit motifs through use of position weight matrices [2].

BioProspector. Also a widely popular tool, BioProspector provides an alternative algorithmic approach to MEME. Although both tools use position weight matrices in their respective algorithms for motif discovery, BioProspector uses a version of Gibbs sampling which is a randomized algorithm as opposed to the deterministic EM algorithm used in MEME. This contrast in the flavor of these two tools was a main factor for their inclusion in the pipeline [3].

CMF. The Combinatorial Motif Finder (CMF) is a motif discovery tool unlike the previous tools. It's approach does not involve probabilistic models, but rather exhaustive combinatorial pattern matching [4]. This adds even more variety to the tools utilized by the pipeline.

Weeder. Similar to CMF, Weeder is a tool that uses a word-based approach to discovering motifs. This type of analysis begins with a set of oligomers which is then expanded combinatorially to provide a matched pattern [5].

DECOD. Finally, DECOD is much different than all of the previous tools; it uses a discriminative approach to solving the problem of motif finding. Through deconvolution and heuristic hill climbing, DECOD attempts to search for a position weight matrix that maximizes the target function [6]. The discriminative approach provides a significant difference in the methodology of this tool in comparison to the others, which allows the pipeline to be even more comprehensive.

2.1.2 Combining Results

Once each of the tools has completed its analysis of the dataset, results from each tool are compiled and stored in a file on disk. From there, the compiled results are then used by the next phase of the pipeline which analyzes the similarity between results and reports a list of the most likely motifs. These motifs are determined by a comparison algorithm that uses dynamic programming to find the edit distance between each combination of motif results from each tool. For example, each of the results generated by MEME are then compared to each of the results for CMF, DECOD, etc. but not its own results. In addition, once the comparison algorithm has completed, the pipeline then begins a searching algorithm that finds all occurrences of the most likely motifs in the original dataset (similar to the computation of p-value) and then assigns a score to each of the found motifs.

As an extension of the previous work done by my teammates, I went on to implement an algorithm to compute the binomial p-value. This addition to the pipeline proved very useful for easy comparison of results, as well as providing important statistical

information about the specificity of the pipeline's results. Essentially, most of the algorithm is completed using the `binom_test()` function of the `scipy.stats` package. The parameters $\theta = \{x, n, p\}$ for `binom_test()` are calculated by going back through the dataset and recording various statistics. Parameter x is the number of distinct sequences in which an occurrence of motif m is found, n is the total number of sequences in the dataset, and p is estimated using the sample mean $\bar{X}$ of motif occurrences per sequence as an unbiased estimator for the probability of an occurrence of motif m existing in an arbitrary sequence[2],

$$\hat{p} = \bar{X} = \frac{M}{N}$$

where M is the total number of occurrences of the motif m and N is the total number of possible locations for the motif m.

2.1.3 Visualization

Weblogo provided the main functionality in our pipeline for creating sequence logos from the resulting position weight matrices generated by our pipeline. Additionally, a simplified motif location diagram is included along with the sequence logo in a generated HTML document which displays the sequence name, the specific motif match, and the location of the motif match within the original sequence.

2.2 Automated Testing Suite

The automated testing portion of this project was another section that I developed heavily. The suite consists of two parts, a synthetic data testing script as well as a real data testing script. Both scripts have similar structure, automatically running the pipeline on a large number of datasets; however, the synthetic data testing script generates its own datasets in addition to running them through the pipeline. The goal of the synthetic data testing is to find various pathologies in the pipeline's motif discovery abilities. These pathologies can then be analyzed to improve sections of the implementation of the pipeline. On the other hand, the real data testing script was designed to run the same datasets through the pipeline with varying input parameters which are then analyzed to determine which of the tests returned the highest scoring results.

3 Results

Unfortunately, acquiring results for this project required a great deal of time due to the long run-time of some of the tools included in our pipeline. Some of the tests were not completed as adequately as intended, but still provided intriguing results nonetheless.

[2] refer to Appendix 6.2 for additional notes on p-value computation

3.1 Synthetic Data

The first testing that was conducted was completed using synthetic data. The benefit of using the synthetic data was the possibility to fully automate the entire process; the creation of sequence files, the processing of the sequence files through the pipeline, and the analysis of results were all part of the automation. This was made possible because the sequence generator script that I created specifies the embedded motif in every generated sequence inside the sequence name. Therefore, this allows the automated testing script to compare the original embedded motif with the motifs discovered by the pipeline and automatically determine the accuracy.

For most of the synthetic data testing, the pipeline performed fairly well on most of the regular data that was generated. Many of the datasets that contained proportionally large amounts of SNPs within the embedded motifs were not easily discovered by the pipeline, if at all. However, this should not really be a factor for real data because the transcription binding sites generally cannot have a significant proportion of SNPs because then they would not function properly.

3.2 Real Data

Many of the tests included in the automatic testing suite yielded at least some information about the transcription binding site that we are searching for in this project. The testing of the real data was separated into two portions: testing by species and testing by gene. The testing by species produced a surprising lack of results. The trials resulted in found motifs exhibiting p-values consistently greater than the threshold for rejecting the null hypothesis, H_0 : a regulatory motif is not shared between genes of the same species. $1.7014\text{x}10^{-3}$ was the minimum p-value of any of the motifs found between genes of the same species. Therefore, it appears that the genes within each species do not share similar transcription binding sites as the tests failed to reject the null hypothesis H_0.

In the case of testing by gene, there was an abundance of results. In particular, the genes Mad1 and Bub1 frequently exhibited significant p-values. In our project, we deemed $p < 1.0\text{x}10^{-8}$ to be significant. Furthermore, many of the other genes also exhibited significant p-values. Figure 3.1 shows a table with some of the most significant results. As is evident from the results, with a degree of statistical evidence we can reject the null hypothesis for this test, H_0 : a regulatory motif is not shared between the same gene of the different species. As seen in Figure 3.1, occasionally the pipeline will generate the same motif multiple times. This slight error is due to the behavior of some of the underlying tools included in our pipeline that behave this way, causing the pipeline to behave similarly.

Figure 3.1: Testing By Gene, Top 30 Results

Gene	**Motif**	***p*-value**
bub1	GACGGTAG	2.8346223243e-13
bub1	ACAATGACAGC	1.28665739988e-12
mad2	ATACTGAA	1.61249282741e-12
mph1	TCACGATTA	5.97338823939e-12
mad1	GAGAAGAAAT	6.47760684159e-12
mad1	GAGAAGAAAT	6.47760684159e-12
mph1	GAAGGC	2.99318174477e-11
mad2	CGTAAAA	3.35877165242e-11
mad1	ATACTGAG	4.89041248448e-11
mad1	GGGAGAAGG	5.22601892836e-11
mad1	GGGAGAAGG	5.22601892836e-11
bub1	CGGCGTA	1.13390774493e-10
bub1	CAGCAT	1.18411106656e-10
bub1	AACTTGGC	1.28175276664e-10
bub1	AACTTGGC	1.28175276664e-10
mad1	CGCGACACC	1.43582932609e-10
mad2	GTATAC	1.49856434069e-10
bub1	TACAGAATTAG	1.72391535153e-10
bub1	ACATAC	1.95853365155e-10
mad2	ATCGTAAAA	2.13317035947e-10
mad1	CTTAGAGTG	2.88290254004e-10
mad1	CTCAGTAT	3.33559928239e-10
mad1	CTCAGTAT	3.33559928239e-10
mph1	GTCATCACA	3.83885613747e-10
mph1	ACTAGT	4.64282808316e-10
mad2	AAGCCACAT	5.45919720186e-10
mad2	AATTCAACATAAA	6.46633638405e-10
mad2	TTCAACATAA	6.77271115109e-10
mph1	TCGTAT	7.35507676409e-10
bub3	AGATATC	8.393637016e-10

4 Discussion

Overall, the GASP project was moderately successful. The pipeline generated statistically significant results, which could potentially be tested further experimentally. Also, the pipeline could even be improved to offer even more accurate results. Additional analysis features such as clustering could assist greatly in improving upon the current pipeline architecture. Furthermore, modifications and extensions to the scoring mechanism such as adding position-based scoring weight or incorporating q-value to compensate for the large number of input sequences could also benefit the pipeline immensely.

5 Bibliography

[1] Rudner A, Murray AW: The spindle assembly checkpoint.
http://www.med.uottawa.ca/rudnerlab/Adam_Rudner_Lab/Publications_files/Curr%20Opin%20Cell%20Biol%201996%20Rudner.pdf

[2] Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization, Timothy L. Bailey and Charles Elkan. Machine Learning 21, 1995, 51-80.

[3] Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput. 2001;:127-38.

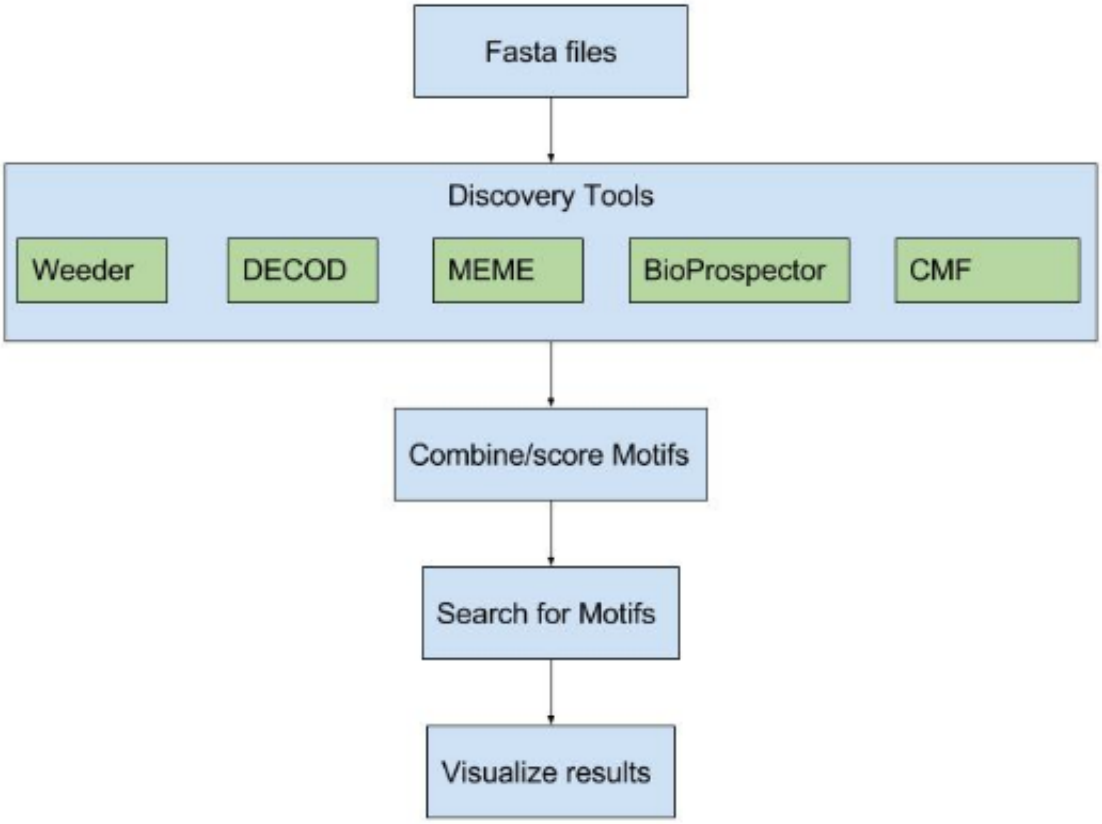
[4] Mason M, Plath K and Zhou Q. (2010). Identification of context-dependent motifs by contrasting ChIP binding data. Bioinformatics, 26: 2826-2832.

[5] Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Giulio Pavesi, Paolo Mereghetti, Giancarlo Mauri2 and Graziano Pesole (2004)

[6] DECOD: Fast and Accurate Discriminative DNA Motif Finding, Peter Huggins, Shan Zhong, Idit Shiff, Rachel Beckerman, Oleg Laptenko, Carol Prives, Marcel H. Schulz, Itamar Simon, and Ziv Bar-Joseph. Bioinformatics 27, 2011, 2361-2367.

6 Appendices

6.1 Pipeline Diagram



6.2 Notes on p-value Computation

This portion of the report took me an extensive amount of time to figure out. I do believe it to be at least a moderately accurate estimate of p-value; however, I am not certain it is not erroneous.