

# Computational Investigation of Gene Regulatory Elements

**Ryan Weddle**

*Computational Biosciences Program  
Arizona State University*

**Jeff Touchman**

*Internship Advisor  
Translational Genomics Research Institute  
Tempe, AZ*

Internship  
Spring 2003 - Spring 2004

Report Number: 04-20

## CONTENTS

|            |                    |    |
|------------|--------------------|----|
| <b>I</b>   | ABSTRACT .....     | 2  |
| <b>II</b>  | INTRODUCTION ..... | 3  |
| <b>III</b> | GOALS .....        | 4  |
| <b>IV</b>  | METHODS .....      | 6  |
| <b>V</b>   | RESULTS .....      | 13 |
| <b>VI</b>  | DISCUSSION .....   | 16 |

## APPENDICES

|            |                  |    |
|------------|------------------|----|
| <b>I</b>   | APPENDIX A ..... | 20 |
| <b>II</b>  | APPENDIX B ..... | 21 |
| <b>III</b> | APPENDIX C ..... | 22 |

## REFERENCES

|           |                             |    |
|-----------|-----------------------------|----|
| <b>I</b>  | ACKNOWLEDGEMENTS .....      | 32 |
| <b>II</b> | SELECTED BIBLIOGRAPHY ..... | 33 |

## ABSTRACT

Laser capture micro-dissection and micro-array analysis now allow the identification of specific mRNAs that are differentially expressed between specific cell populations. Laser capture micro-dissection makes it possible to obtain samples of different cell types, even from within the same tissue. These technologies have been used to identify mRNAs distinguishing between invasive and non-invasive glioma cells. In this project we leverage available genomic sequence to search for DNA sequence elements that could explain the apparent co-regulation of these genes.

## INTRODUCTION

High grade invasive gliomas are a particularly devastating form of cancer. Glioma is a type of cancer that develops from glial cells in the brain, which then develop into tumors. The cells that form the core of the tumors divide rapidly, but move very little. Cells around the periphery of the tumors tend to be motile, exhibiting a different phenotype that allows them to move outward into the brain. While traditional cancer therapies are occasionally effective at combating the tumors themselves, they are less effective at preventing recurrence due to the invasion of the motile cells into other areas of the brain. It is hoped that by understanding the mechanisms which cause some tumor cells to become motile, more effective therapies might eventually be developed.

Laser capture micro-dissection has been used to isolate these two populations of cells in order to perform micro-array analysis on them. The results of micro-array studies performed at Tgen (Neurogenomics) have shown that there are fifteen genes, which appear to be differentially expressed between the two populations. These findings were further verified by quantitative PCR also performed at Tgen.

It is hoped that by understanding the mechanisms regulating the expression of these genes, better treatments could be discovered for dealing with invasive glioma cells. We hope to use computational methods to identify the mechanisms which regulate this set of genes. In order to understand the regulation of these specific genes, however, it is first necessary to understand something about eukaryotic gene regulation in general.

Gene regulation occurs at four levels within the cell. At one end, the process of transcription itself is regulated by a variety of mechanisms – with higher rates of transcription obviously leading to higher levels of transcript in the cell. At the next level, abundance of the transcript is regulated by modifications affecting their rate of degradation. Translation is then directly regulated by both message specific and more general means such as relative amino acid

abundance. Lastly, post-translational modifications ultimately determine the amount of a protein that is active within the cell.

We are primarily interested with regulation of the transcription process itself. In order to understand how co-regulation can explain the correlation of transcript levels, it is necessary to understand how the process of eukaryotic transcription. Transcription, simply put, is the process that copies DNA sequences into RNA. In humans, all protein-coding genes are transcribed by RNA Polymerase II. Unlike bacterial RNA Polymerases, eukaryotic RNA Pol II class genes are typically under positive control – they are not transcribed unless certain specific conditions are met.

Before any eukaryotic gene can be transcribed, the chromatin structure of the DNA must be open in the region of the gene. Opening the chromatin may involve modification or movement of the histone octamers, but it does not typically involve their removal. Once the chromatin is opened, however, the RNA polymerase is still unable to bind directly to the promoter regions of the DNA. It is first necessary for a special DNA binding protein called a transcription factor to bind to the DNA. The RNA polymerase is then able to bind directly to this transcription factor in order to attach itself to the DNA in preparation for transcription.

Additional transcription factors also serve to enhance or suppress transcription of the gene. They bind to specific, but variable, sequence sites, which may occur within the gene itself or many thousands of base pairs away. Most frequently they bind to the region immediately upstream of the gene or within one of the introns of the gene. In typical eukaryote genes, transcription is finely regulated by the competing actions of many of these DNA binding proteins. When many such transcription factors act in concert to precisely control gene expression, they are often referred to in aggregate as a regulatory cassette. The immediate upstream region which facilitates binding of the polymerase is frequently referred to as the promoter region of the gene.

## GOALS

The goal of this project is to use computational methods of genome sequence analysis in an attempt to understand how the differentially expressed genes are regulated. All of these methods are based on the fact that transcription factor binding sites, though highly variable, can be statistically characterized and modeled through a variety of means. We hope to apply a number of techniques in order to identify putative binding sites that may be involved in the regulation of these genes.

Implicit in this goal are a number of assumptions. First, we are assuming that these genes, if co-regulated, are in fact regulated by the same transcription factors. This is in contrast to the situation where the genes are regulated by different transcription factors, all controlled by a common mechanism somewhere upstream in the pathway. Additionally, we assume that the binding sites of these transcription factors will exhibit sufficient similarity across sequences to be identifiable. Lastly, any analysis is predicated on the assumption that the regulatory regions are in some way local to these genes – again, this is not always the case, as binding sites and promoters many thousands of base pairs away can affect the transcription of a gene.

As this is an exploratory investigation our goal is two-fold. First, we hope to evaluate a variety of known methods in searching for transcription factor binding sites in our sequence data. Second, we hope to form a testable hypothesis so that we can say something about the statistical certainty of any findings. The first goal is ultimately to suggest avenues for further investigation, while the second goal is to determine if pursuit of that further investigation is justified. Ideally, we hope to identify those sequence elements, which we think exert regulatory control over this set of genes. If this proves impossible, we hope to at least characterize the sequence similarity within this set of genes in comparison to sequence similarity between other genes at random, based on the hypothesis that such sequence similarity could be indicative of common regulatory elements.

## METHODS

Investigation of our research goals required the use of a number of different computational methodologies. Phylogenetic footprinting, repeat sequence masking, transcription factor binding site detection, motif discovery and detection, association rule mining, as well as novel short sequence methods were used. Perl was used extensively to automate the analyses and to convert data between a variety of file formats.

### Phylogenetic Footprinting

---

All methods employed have in common the use of phylogenetic footprinting. Phylogenetic footprinting refers to the process of aligning genomic data from two relatively closely related species in order to determine what regions of the genome have been conserved. The principle assumption of phylogenetic footprinting is that “function is conserved”. In other words, functional sequence elements are conserved with fewer base changes than sequence elements serving less specific functions. The most highly conserved sequence regions are those that code for protein sequences of genes. Conservation of introns and genomic DNA is markedly lower than that of codons. There are, however, highly conserved regions of non-coding DNA. To the extent that regulatory sequence elements are conserved between closely related species, it is expected that they exist within these conserved non-coding regions.

Due to limited availability of mammalian genomes when this project was initiated, footprinting was done against only one other genome. *Homo sapiens* sequence data was aligned with *mus musculus* sequence data in order to identify the regions of conservation to search for regulatory elements. After trying several alternatives, it was decided that the *blastz* (Schwartz et alia, 2003) component of the *Pipmaker* (Schwartz et alia, 2000) software package would be used to do the alignments.

### Repeat Sequence Masking

---

Phylogenetic footprinting was used, in a sense, to reduce the search

space by limiting the sequence regions that needed to be considered. Repeat sequence masking served a similar purpose, but in addition to search space reduction, it also helped to eliminate “false positive” like discoveries. To this end, the Repeatmasker (Smit et alia) software package was used. As this was more of an investigation than a single large scale experiment, it could not be certain that any results found were actually regulatory in nature without biological verification. If any repeat sequences occurred more than once in the genomic sequence data, they would have likely been found by our pattern mining. Since repeat sequences have higher similarity across occurrences than typical regulatory elements, such as transcription factor binding sites, their presence in the sequence data could actually overshadow the signal from other potentially interesting patterns.

### **Transcription Factor Binding Site Detection**

---

Transcription factors are DNA binding proteins that perform a number of roles facilitating and regulating the transcription of genes. In eukaryotes, RNA Polymerase is not able to bind directly to the DNA, and thus requires a transcription factor to bind and provide a platform for it to associate. Other transcription factors bind to enhance or suppress expression, or to facilitate other necessary events such as opening the chromatin structure around a gene. These DNA binding proteins are sequence specific, but not strictly so. The loose specificity has at least two identifiable causes. First, DNA binding proteins are not so much recognizing the individual bases themselves, but the chemical environment created by the sequence of bases. Second, transcription factor bind sites exist in stronger and weaker varieties, due to slight conformational changes in the DNA structure caused by sequence variation. Different sequence variations create slightly different binding potentials at the sites, thus different sequences have evolved to bind the same proteins with different strengths.

Due to the high degree of sequence variation in binding sites for a specific transcription factor, consensus sequence representations are not

sufficient for identifying them. Instead, they are represented by weight array methods (WAM) (Zhang, Marr, 1993) or weight matrix methods (WMM). The Transfac database catalogues these representations, along with consensus sequences, for already characterized transcription factor binding sites. We used MatInspector as well as the raw Transfac (Matys et alia, 2003) database to search for these sites.

We wrote a Perl program to use the Transfac matching software to automatically generate an Excel workbook summarizing the distribution of known transcription factor matches in the conserved sequences. Different sheets were automatically inserted to provide different views of the data under varying parameters. The intent of this was to provide a platform for the application of varying data mining algorithms. In practice, there was little to be gained from the application of clustering algorithms to this data.

## **Motif Discovery and Detection**

---

The Transfac database is useful for the detection of those transcription factor binding sites that have already been described. Even though Transfac can be used to extrapolate and discover new sequences that are good candidates for binding sites of known transcription factors, it cannot be used to find sites that do not correspond to an already characterized transcription factor. In order to search for novel binding sequences it was necessary to use pattern detection software.

We experimented with the use of the MEME/MAST software package from the UCSD Department of Computer Science and the San Diego Supercomputing Center. MEME stands for Multiple Em for Motif Elicitation (Bailey, Elkan, 1994) while MAST (Bailey, Gribskov, 1998) stands for Motif Alignment and Search Tool. MEME works through fitting a mixture model of multiple expectation maximization calculations to find motifs repeated in DNA sequence data. MAST is used to search for other occurrences of motifs identified using MEME. We used these methods to search the conserved regions of our genes for new patterns.

## Association Rule Mining

---

As transcription factors act in concert to regulate the expression of a gene, it would be useful to see whether any patterns in their occurrence could be discerned. One means used to find patterns that have low support but high confidence is association rule mining. Association rule mining (ARM) (Agrawal, Srikant, 1994) is an algorithm which finds frequent itemsets and association rules. A frequent itemset is defined as a collection of items, which co-occur frequently. In our case, this would be either a set of transcription factors which tend to occur on the same genes, or it could be a set of genes which tend to have the same transcription factor binding sites. In some domains it is clear in which way the problem should be framed, but in our domain a case can be made for either way. An association rule is a rule that states that whenever one particular frequent itemset occurs, another itemset occurs with a very high probability. In our domain, this would be a rule indicating that whenever a certain set of transcription factors occur associated with a particular gene.

The main method for discovery of association rules is known as the apriori algorithm. We used the ARMiner software to search for association rules in our data set, framing the problem in both ways described above.

## Novel Short Sequence Methods

---

Investigation into association rule mining prompted us to develop novel short sequence methods for genome similarity analysis. All of our analyses so far had dealt with some subset of the conserved sequence region – typically those regions found by some variety of pattern matching (or pattern discovery). Both Transfac and MEME/MAST analyses ultimately left us with these shortcomings. We were interested in a means to analyze the totality of short sequences represented by the conserved areas of our target genomic regions. Since association rule mining scales to large datasets very well, our initial plan was to frame the problem in such a way that we could search for association rules over arbitrary sequence elements. Furthermore, with modifications, association rule mining can find rules with low frequency, but high support.

Considering all subsequences of a set of genes would be computationally expensive in terms of both memory and time. Because of this, we decided to focus on sequences of lengths likely to correspond to transcription factor binding sites. Based on the Transfac core sequences, we decided to focus on sequences in the range between 6bp and 11bp. A database of all such kmer (k = 6 to 11) sequences occurring more than once in our target genes was generated by a Perl program written for this express purpose. In addition, we wanted to allow for a small amount of mismatch, so all near matching (up to several bp differing) matches were calculated as well. We found what seemed to be a surprisingly high number of exact matches of lengths all the way up to 11bp. In fact, there were 73 exactly matching 11mers occurring in two or more genes. On this basis, we were able to further refine our the focus of our research:

**Is there more similarity between our target genome regions than would be expected between randomly selected genome regions?**

In order to answer this question, we needed to address just what was meant by “randomly selected genome regions.” In order to be certain that our target regions have more 11mer similarity than would be expected due to chance, we would need to run a number of trials to perform a statistical hypothesis test. Each trial would require randomly selecting a set of thirteen genes (the same number in our target set) and repeating our analyses on them. This was complicated by the fact that our analyses were not, at this point, entirely automated. The most time consuming step by far was the actual acquisition of homologous human and mouse genome regions for each gene of interest, which had been downloaded from NCBI by hand for our target gene group. We addressed this issue by initially using randomly simulated genomic data, and then, having verified that it was worthwhile, devising an automated procedure to randomly select genes from the entire set of human-mouse homologs, extract the appropriate sequence regions, and perform the analyses on them.

As mentioned above, we initially tested our hypothesis by generating random genomic sequence data for our trials. The entire process was handled by a series of Perl scripts. For each trial, the programs calculated the base composition of each conserved sequence fragment in our target data set. New fragments were randomly generated which exactly preserved the base composition of the original sequence data. The 11mer analysis was then performed on the corpus of new sequences generated by this process. One hundred random trials were performed in this manner. This process was repeated with a uniform nucleotide distribution for further comparison.

Encouraged by the first test, we decided it was necessary to more concretely test the hypothesis. The first test depended on randomly generated genomic sequence data, but in reality it would be more correct to use randomly selected genomic sequence data. In order to do this, it was necessary to automate the process of genomic sequence extraction and homolog detection. Fortunately, we found it was possible to automatically download all of the human-mouse homolog genes from EMBL. A Perl program was then written to perform all of our pre-processing (footprinting, sequence extraction, etc.) on this whole genome corpus, resulting in a sequence database that could be used for randomized 11mer analysis. Again, this process was very data intensive, using several gigabytes of filespace on disk.

Once we had a pre-processed database, it was relatively simple to write a Perl program to randomly select sets of genes and perform the 11mer count analysis on them. This set of scripts randomly selected sets of thirteen genes and ran the same 11mer analysis on them that we had used for our target set (see Appendix C). One hundred trials were performed and the counts were recorded for each trial, along with the average.

In addition to the comparative analyses of 11mers, we also applied several additional methods to investigate the 11mers that we did find in our target gene set. We statistically evaluated positional information regarding where each 11mer occurred within the genomic region associated with its gene.

Additionally, we applied some of our the methods previously discussed to analyze just the 11mer sequences themselves in order to search for any significant patterns.

## RESULTS

As an exploratory investigation, this project can be seen as having had two sets of results. On the one hand, we successfully used and evaluated a number of means of analyzing our target genes to look for patterns that could explain their apparent co-regulation. Additionally, we were able to use some of these methods to investigate our research hypothesis – that our putatively co-regulated genes could be discriminated from some other random set of genes on the basis of the presence of some identifiable sequence element.

### Exploratory Investigation Results

All of our methods employed phylogenetic footprinting to some degree or another. Phylogenetic footprinting was very effective at reducing the size of the search space by upwards of seventy five percent. While it is possible that there could exist regulatory elements in the human genes that have not been highly conserved over time, and it is in fact likely that they have undergone some variation, it would be surprising if such elements were not more highly conserved than the background sequence. Phylogenetic footprinting between mouse and human resulted in a corpus of 792 regions of ungapped alignment in our target set of genes.

Masking out coding and repeat sequences was also effective in reducing the search space, as well as in eliminating non-target signal from the sequence. By removing these sequences, spurious non-regulatory results, which would otherwise overwhelm any regulatory sequences were avoided.

We successfully utilized the Transfac database to find matches for known transcription factors in the conserved sequences. The application of data mining algorithms to the resulting data did not provide particularly significant results – there were too few genes for clustering over transcription factor occurrence to provide interesting or meaningful results. Depending on the parameters used association rule mining over known transcription factor binding sites found either very few or very many rules. Unfortunately, mining rules directly

on transcription factor matches frequently found spurious rules due to two similar or overlapping binding sites. MEME based pattern discovery found very few motifs, the only one occurring with a significant p-value turned out to be a previously characterized splice site, rather than a novel motif.

## **Research Hypothesis Results**

---

We investigated our research hypothesis, that “there is more similarity between our target genome regions than would be expected between randomly selected genome regions,” through the kmer analysis as mentioned in the methods section. A series of databases were created in order to facilitate this investigation. The first kmer database generated, including all 6-11bp kmers represented in our target gene set, resulted in roughly 150,000 occurrences of approximately 80,000 unique kmers. This database allowed for mismatches of up to several base pairs, depending upon the kmer length. On disk it occupied 550MB of space. Removing all inexact matches from the database still resulted in a database of 40MB. A smaller database considering only the 792 ungapped alignments resulting from phylogenetic footprinting over our target data set resulted in a count of 73 exactly matching 11mers.

When this analysis was repeated over 100 sets of randomly generated sequence data, we found that there were, on average, 36.28 exactly matching 11mers per set. The standard deviation of number of exactly matching 11mers was found to be 7.9. These results (see Appendix A) provided an initial confirmation of our research hypothesis:

In a one sided z-test having null hypothesis that the population mean corresponding to the random experiments is greater than or equal to 73, and having the alternative hypothesis that the population mean is less than 73, a z-score of -46 was obtained. Such a high test statistic corresponds to a p-value of less than  $10^{-6}$  – at any level of certainty we would reject the null hypothesis and accept the alternate hypothesis that the random population mean is lower than the 73 occurrences observed in our experimental data. This provided a strong indication that the sequence similarity in our target group was greater than

would be expected by random chance.

Although our results confirmed that the sequence similarity in our target group was greater than would be expected by random chance, our results did not confirm that the sequence similarity is greater than would be expected within the human and mouse genomes in general. When we repeated our experiment with randomly selected genes (see Appendix B), rather than randomly generated genomic sequence data, we found that the number of 11mers within our target group was actually lower than the population average of 348.11. In this case, we clearly fail to reject the null hypothesis that the population mean number of 11mers is greater than or equal to 73. In other words, there is insufficient evidence to conclude that our target set of genes shows more short sequence similarity than randomly selected sets of genes from the universe of mouse human homologs.

The non-comparative investigation of the set of 11mers also yielded some interesting, if anecdotal, results. A slight positional correlation was observed in the occurrence of 11mers in our target gene data set. The regions including the 2000bp of genomic data immediately upstream from the beginning of the gene included 31 of the 73 perfect matching 11mers. Since we considered the 10,000 bp upstream and the introns, this immediate upstream region is clearly over-represented in terms of kmer occurrence. We additionally noted that 41 of the 73 perfect matching 11mers had each of their matching sites occurring within 1000bp of each other – when each was measured relative to the beginning of the upstream region it occurred in. In addition to this positional information, the search against the Transfac database discovered a number of known transcription factor binding models which had good matches on the 11mers. Of particular interest were the matches for TATA boxes and promoter cap sites.

## DISCUSSION

### Exploratory Investigation

---

As an exploratory investigation, this research was quite successful. We were able to evaluate the utility of a number of methods of searching for regulatory elements. Additionally, we were able to develop a research hypothesis and clearly test it against control data. There were, however, a number of impediments that made our task more difficult.

First and foremost, there were problems due to the exploratory nature of this research. We were attempting to investigate a specific novel data set, while at the same time researching the methods necessary to do so. While there are many models of gene regulation in lower eukaryotes such as yeast, the biology of gene regulation in higher eukaryotes is much more complicated and is still being actively investigated. This made it difficult to establish any control group to evaluate our investigations against. As a result, we ended up framing the experiment in terms of sequence similarity, which, although likely to bear on regulatory mechanisms, does not guarantee their presence by any means.

There are also a number of confounding biological factors that we were unable to control for or investigate. Our study focused on the regions 10,000bp upstream of the genes, as well as on the introns. Quite frequently, however, enhancer sites in eukaryotes can be much further upstream (or downstream) than this and can act on a number of genes in their vicinities. These enhancers act when the DNA strands fold over in such a way that they are able to contact promoters or enhancers near or within the individual genes. Current methods are unable to accurately predict these kinds of conformational changes, and we were thus unable to take them into account.

Another structural factor integral to the understanding of gene regulation is the structuring of the chromatin around the genes of interest. Chromatin

structure cannot be wholly predicted from sequence information, but is directly involved in the regulation of expression in eukaryotes. Although transcription factors are often involved in the opening of the chromatin in a specific region, there are many elements involved in regulating this process. The specific pattern of open and closed chromatin appears to be epigenetically inherited in individual tissues and cell lines. Patterns of chromatin structure, then, can explain the apparent co-regulation of genes, which may not be activated by common pathways. Moreover, even genes turned on in the same pathways may have their expression activated by different downstream transcription factors.

Clearly, then, there are a number of biological issues at play, which are very difficult to adequately take into account. Combined with the limited data available, it is unsurprising that we were unable to establish very many conclusive results. Further studies would benefit from either focusing on proven methods or using larger, better characterized data sets.

## **Research Hypothesis**

---

On final analysis, we are unable to conclusively affirm our research hypothesis, that there is significantly more 11mer sequence similarity within our target gene set than would be expected at random. Our initial investigation of this question appeared to strongly support this assertion, but our subsequent, more exhaustive, investigation appeared to undermine that conclusion. Our non-comparative investigations, however, seem to support the notion that the identified short sequences may be regulatory in nature. Nonetheless, the current statistical evidence is insufficient to support the conclusion that our target set of genes differ significantly in their short sequence similarity than sets of genes selected at random. At this point, we cannot assert to have uncovered any regulatory mechanisms.

Nonetheless, we must attempt to understand the differences observed between the experiments involving the randomly generated sequence data, and the experiments involving the randomly selected genes. Beyond the issue of generation versus selection, there are significant differences between these two

experiments. In the case of the randomly generated data, the only additional processing necessary was the kmer analysis. Each kmer analysis in the random trials, used a sequence corpus of identical total length. In contrast, phylogenetic footprinting had to be performed for each random trial in the selection analysis, resulting in sequence corpuses of different length for each trial. Assuming a fixed apriori probability of kmer matches, there would then be variability in kmer match frequency with the size of the sequence corpus.

There are two main factors that could affect the size of the sequence corpus used in the kmer analysis. First, since we include 10,000bp upstream and all introns, the size of the introns could significantly affect the amount of sequence ultimately utilized in the kmer analysis. Second, the rate of divergence of the genes would affect the proportion of the sequence conserved as ungapped alignments. Housekeeping genes, for instance, may be more highly conserved than other more specialized genes. If this were true of the genomic data in the region around those genes as well, then we would expect kmer analyses of housekeeping genes to result in more matches. Thus, if the genes of our target set have undergone more drift than the average gene, then this might be one explanation for their having fewer kmer matches than other randomly selected genes.

### **Suggestions for Future Work**

---

While failing to conclusively validate any methods, this investigation recommends a number of avenues for further study. As was suggested in the discussion of the research hypothesis results, the random selection experiment could be repeated with a number of modifications to control for different conserved sequence lengths due to differing gene sizes and rates of divergence. It would also be highly desirable to obtain a better characterized data set to use for investigation of novel computational methods for discovering regulatory mechanisms.

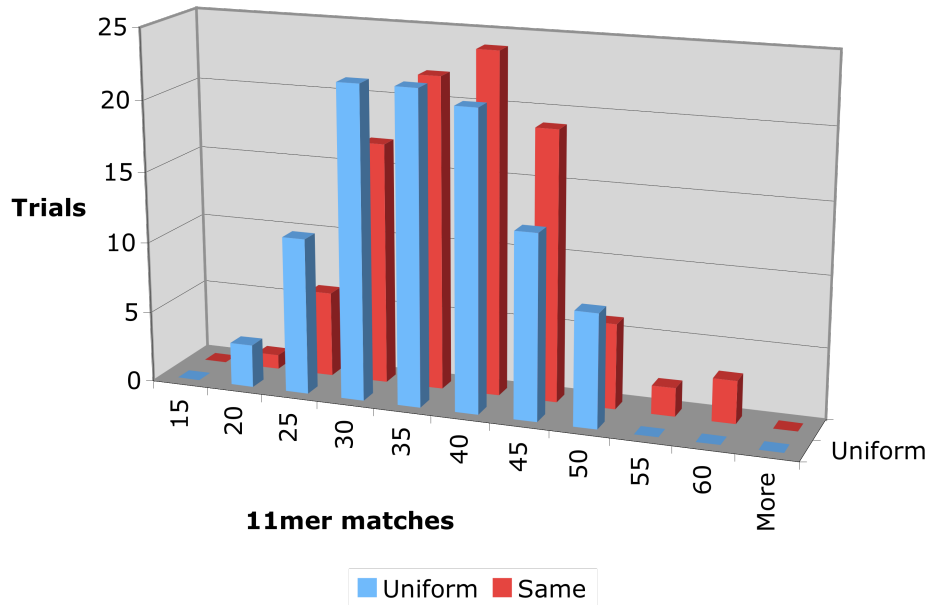
If a better data set were available for similar investigation, the technique of phylogenetic shadowing would also prove highly beneficial. Phylogenetic

footprinting uses the difference between two DNA sequences to determine what regions have been under greater selective pressure. Two very closely related species will result in very high conservation, as there has not been adequate time for genetic drift in the regions not under pressure. Two distantly related species, on the other hand, may show little conservation of interest beyond the coding sequences themselves due to evolution of regulatory mechanisms over evolutionary time. Phylogenetic shadowing, on the other hand, would let us analyze the small differences between a large number of very closely related species in order to get a better picture of the selective pressure facing the genomic sequence. This would essentially serve to improve the precision and resolution of the analyses for identifying functional conserved sequence regions.

# APPENDIX A: RANDOMIZED SEQUENCE TRIAL RESULTS

## Histogram of 11mer Count Distributions

**11mer Distributions, Uniform and Controlled**



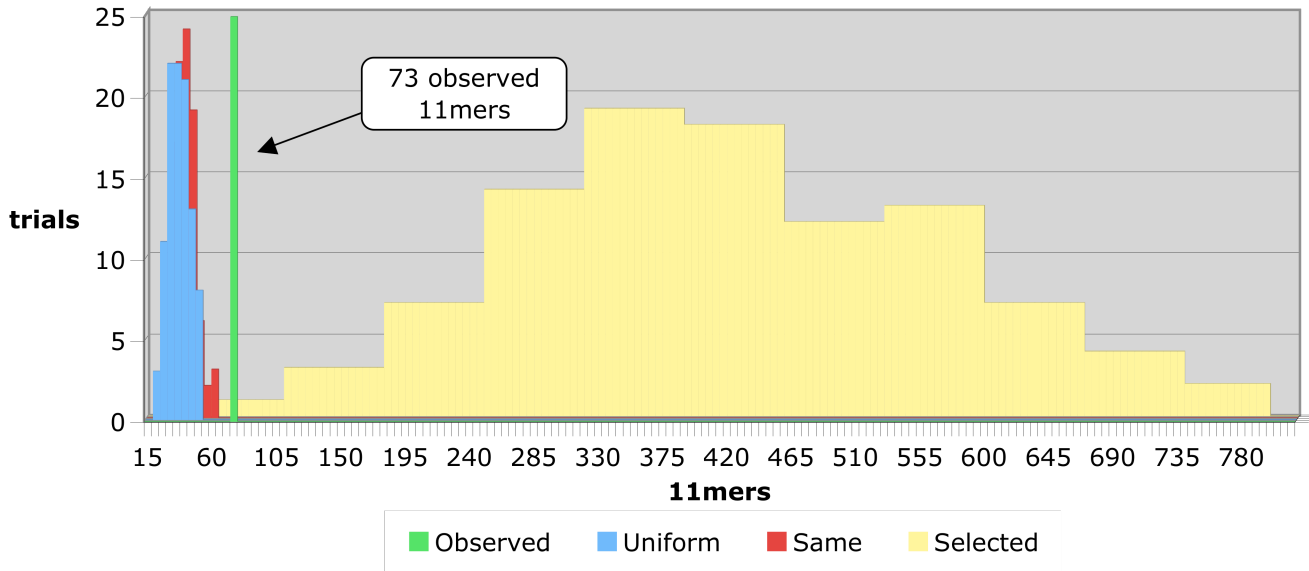
One hundred random trials of 11mer analysis were performed under two distributions: the uniform nucleotide probability distribution and a distribution mirroring nucleotide bias from our target gene set ( A=26.7% C=23.0% G=26.1% T=24.2% ). Histograms were constructed showing the distribution of 11mer counts in the random trials.

## Summary Statistics

| 11mers    | Uniform | Controlled | Z-test  |
|-----------|---------|------------|---|
| Mean      | 33.65   | 36.28      | <p>Is our observed count of 73 11mers higher than the controlled population mean?</p> <p style="text-align: center;">Z-score = -46<br/>P-value = 10<sup>-6</sup></p> <p>Conclusion: At any reasonable level of certainty, our observed score is higher than the controlled population mean.</p> |
| Median    | 33.50   | 36.00      |   |
| Variance  | 57.89   | 62.51      |   |
| Std. Dev. | 7.61    | 7.91       |   |
| Min       | 16.00   | 18.00      |   |
| Max       | 49.00   | 58.00      |   |

# APPENDIX B: RANDOMLY SELECTED GENE SET TRIAL RESULTS

## Superimposition of Histograms for Three Distributions



All three distributions are represented here by superimposed histograms. The observed 73 11mer matches are indicated by the vertical stripe at 73. The two random sequence distributions are compressed here to the far left. The large distribution reflects the distribution of 11mers found by analysis of randomly selected sets of mouse-human homologs.

## Summary Statistics

| 11mers    | Uniform | Controlled | Selected  | Comparison   |
|-----------|---------|------------|-----------|--|
| Mean      | 33.65   | 36.28      | 348.11    | It can be easily seen that our experimental observation of 73 11mers is less than the randomly selected population mean of 348.11. |
| Median    | 33.50   | 36.00      | 336.50    |  |
| Variance  | 57.89   | 62.51      | 22,628.02 |  |
| Std. Dev. | 7.61    | 7.91       | 150.43    |  |
| Min       | 16.00   | 18.00      | 35        |  |
| Max       | 49.00   | 58.00      | 738       |  |

## APPENDIX C: MAIN KMER CODE

### Randomtrial.pl

---

This script performs a number of trials specified on the command line over a randomly selected set of genes. It outputs the result for each set of genes on a newline and prints the average on the last line of output.

```
#!/usr/local/bin/perl -w

$num_trials = shift;
$sum=0;
$result=0;
for ($i=0;$i<$num_trials;$i++) {
    $result = `./pickRandomGenes.pl 13 | grep -c ">"`;
    print "$result";
    $sum+=$result;
}
print $sum/$num_trials;

print "\n";
```

### pickRandomGenes.pl

---

This script reads a number of trials from the command line and selects that many genes from at random from the genome database in REGIONS\_DIR. It then uses the script build\_11mers.pl to find the common 11mers common to the randomly selected set of genes.

```
#!/usr/local/bin/perl -w

use strict;

my $num_trials = shift;
my $REGIONS_DIR = "regionsSingleStrand";

my @regionList = glob("$REGIONS_DIR/*.fasta");
my $num_genes = scalar(@regionList);
```

```

my $regionIndex;
my $results;
my $selectedGenes = "";

#for (my $i=0;$i<$num_genes;$i++) {
#  $results += `grep -c ">" $regionList[$i]`;
#}
#$results /= $num_genes;
#print "$results\n";

for (my $i=0;$i<$num_trials;$i++) {
  $regionIndex = $num_genes*rand();
  $selectedGenes .= "$regionList[$regionIndex] ";
}

$results = `cat $selectedGenes | build_11mers.pl`;
#$results = `cat $selectedGenes | make_kmers.pl`;
print "$results\n";

```

## **build\_11mers.pl**

---

This program takes a list of genes and makes a database of the nmers that they have in common. For simplicity in this phase, it has been hard coded to compute only the 11mers, as they are the benchmark being used for comparison. It is used by pickRandomGenes.pl to determine the 11mer content of the randomly selected genes. It was also used manually to count the 11mers in our target dataset of co-regulated genes.

```

#!/usr/bin/perl

my @lines = <>;
my %genes;

my %regions;
my %kmers;

my %rlists;
my %glists;

my %genelookup;

my @lengths = (11);

```

```

main(\@lines);

sub main {

    my $lref = shift @_;
    my @lines = @$lref;

    while(@lines) {
        my $headertext, $sequence;
        my @header;
        my $region;
        my $gene;

        $headertext = shift @lines;
        $sequence = shift @lines;

        @header = parse_header($headertext);
        $region = name_region(\@header);
        $gene = $header[0];

        $regions{$region} = $sequence;
        $genelookup{$region} = $gene;
    }

    foreach (@lengths) {
        my $n = $_;
        add_nmers($n);
    }

    my $i = 0;

    foreach(keys(%kmers)) {
        $i++;
    }

    findpairs();

    #    print "$i kmers found\n";

}

sub findpairs {

    foreach (keys(%glist)) {
        my $kmer = $_;

```

```

my $lref = $glist{$kmer};

my @list = makeunique(@$lref);

if (scalar(@list) > 1) {
    print "> $kmer";
    foreach(@list) {
        print "___";
        print "$_";
    }
    print "\n";
    print "$kmer\n";
}

}

}

sub makeunique {
    my @list = @_;

    my %ht;

    foreach (@list) {
        $ht{$_} = 1;
    }

    my @newlist = keys(%ht);
    return @newlist;
}

sub add_nmers {
    my $n = shift @_;

    foreach(keys(%regions)) {
        add_nmers_from_region($n, $_);
    }
}

sub add_nmers_from_region {
    my $n = shift @_;
    my $region = shift @_;

    my $seq, $start, $end, $seqlength;
    my $gene;

    $gene = $genelookup{$region};

```

```

$seq = $regions{$region};
$seqlength = length($seq);

$start = 0;
$end = $start + $n - 1;

while ($end < $seqlength-1) {
    my $kmer;
    $kmer = substr($seq, $start, $n);
    chomp $kmer;

    $start = $start + 1;
    $end = $start + $n - 1;

    $kmers{$kmer} = 1;

    #print "$kmer\t$region\t$gene\n";

    push @{$glist{$kmer}}, $gene;
    push @{$rlist{$kmer}}, $region;
}

}

sub parse_header {
    my $header = shift @_;
    my $gene, $start, $stop, $pcnt;

    $header =~ m/>\s+(\S+)\s+(\d+)\s+(\d+)\s+(\d+)/;

    $gene = $1;
    $start = $2;
    $stop = $3;
    $pcnt = $4;

    my @header = ($gene, $start, $stop, $pcnt);

    return @header;
}

sub name_region {

    my $href = shift @_;
    my @header = @$href;

    my $gene = $header[0];

```

```

my $number, $name;

if($genes{$gene}) {
    $genes{$gene} = $genes{$gene} + 1;
} else {
    $genes{$gene} = 1;
}

$number = $genes{$gene};

$name = "$gene:$number";

return $name;
}

```

## blastall.pl

---

This script reads pairings of mouse and human genes and uses the blastz program to identify the conserved regions. In practice, it was run over correlatedGeneIDs.txt to footprint all annotated genes in the human genome which have a homologous match in the mouse genome. Cutregions.pl was then run to create the database which is randomly sampled in pickRandomGenes.pl.

```

#!/usr/local/bin/perl -w

my $hdir = "./humanGenes";
my $mdir = "./mouseGenes";
my $bdir = "./blastzGenesSingleStrand";
my $blastz = "/usr/local/genome/bin/blastz";

my $count = 0;

while (my $line = <>) {

    my @elts = split /\t/, $line;
    my $humanGene = $elts[0];
    my $mouseGene = $elts[1];
    chomp $mouseGene;

    my $outputName = "$humanGene.$mouseGene.blastz";

```

```

`$blastz $hdir/$humanGene.fasta $mdir/$mouseGene.fasta >
$bdir/$outputName B=0`;

$count ++;
if ($count%100==0) {print STDERR "$count\n";}

}

```

## cutregions.pl

---

This script processes the results of blastz execution from blastall.pl, and uses them to cut the corresponding regions from the original genome sequences. This creates the database that is randomly sampled in pickRandomGenes.pl.

```

#!/usr/bin/perl -w

my $blastzdir = "./blastzGenesSingleStrand";
my $humandir = "./humanGenes";
my $regionsdir = "./regionsSingleStrand";

main();

sub keep_region {
    my $gene = shift @_;
    my $pcnt = shift @_;
    my $start = shift @_;
    my $end = shift @_;

    if ($end < 10000) {
        return 1;
    } else {
        return 0;
    }
}

}

sub main {

    my @array = glob("$blastzdir/*");

```

```

my $i = 0;

foreach $align (@array) {
    $i++;

    my @path = split '/', $align;
    my $names = pop(@path);

    #print "$names\n";
    my @comps = split /\./, $names;
    my $humanGene = $comps[0] . "." . $comps[1];
    my $mouseGene = $comps[2] . "." . $comps[3];

    open (ALIGNF, $align);
    my @blastlines = <ALIGNF>;
    close ALIGNF;

    my @aligns = grep /^s\s\|/, @blastlines;

    open (SEQF, "$humandir/$humanGene.fasta");
    my @seqlines = <SEQF>;
    close SEQF;

#    my @aligns = `grep '^ |' $align`;

    dumpregions($humanGene,$mouseGene,\@aligns,\@seqlines);
        if($i%100==0) { print "$i\n"; }
    }
}

sub dumpregions {

    my $gene = shift @_;
    my $gmus = shift @_;
    my $aref = shift @_;
    my $sref = shift @_;

    my @alignments = @$aref;
    my @seqlines = @$sref;

    my $align_input = join ", @alignments;
    my $seq_input = join ", @seqlines;

    # Clean up the Sequence data...

```

```

my $cseq = clean_fasta($seq_input);

my @aligns = parse_aligns($align_input);

my @newfasta = split_fasta($gene, \$cseq, \@aligns);

my $filename = "$gene.$gmus.fasta";

open (TEMP, ">$regionsdir/$filename");
foreach(@newfasta) {
    print TEMP "$_\n";
}
close TEMP;
}

sub split_fasta {
    my $gene = shift @_;
    my $sref = shift @_;
    my $aref = shift @_;

    my @aligns = @$aref;

    my @fastas;

    foreach (@aligns) {
        my @alignment = @$_;
        my $percent_id = shift @alignment;
        my $start = shift @alignment;
        my $end = shift @alignment;

        if (keep_region($gene, $percent_id, $start, $end) == 1) {
            my $fasta = fasta_substring($sref, $start, $end);
            $fasta = "> $gene $start $end $percent_id\n$fasta";
            push @fastas, $fasta;
        }
    }
    return @fastas;
}

sub fasta_substring {
    my $sref = shift @_;
    my $start = shift @_;
    my $end = shift @_;
    my $wseq = $$sref;

```

```

#   my $nseq = substr($wseq, $start, ($end - $start + 1));
    my $nseq = substr($wseq, $start - 1, ($end - $start + 1));

    return $nseq;

}

sub parse_aligns {
    my $align_input = shift @_;
    my @lines = split '\n', $align_input;
    my @aligns;

    foreach (@lines) {
        my $line = $_;
        $line =~ m/.*\D*(\d+)\D*(\d+)\D*(\d+)\D*(\d+)\D*(\d+)/;

        my @align = ($5, $1, $3, $2, $4);
        push @aligns, \@align;
    }

    return @aligns;
}

sub clean_fasta {
    my $fasta = shift @_;
    my @lines = split '\n', $fasta;
    my $header = shift @lines;

    my @newlines;

    foreach (@lines) {
        $_ =~ s/\W+//g;
        push @newlines, $_;
    }

    my $cleanseq = join "\n", @newlines;

    return $cleanseq;
}

```

## ACKNOWLEDGEMENTS

Dr. Jeff Touchman  
Director DNA Sequencing Center, Tgen  
Assistant Professor, ASU School of Life Sciences

Dr. Phillip Stafford  
Director Biostatistics Center, Tgen

Dr. Rosemary Renaut  
Professor, ASU Department of Mathematics and Statistics  
Director, ASU Computational Biosciences Program

Dr. Michael Berens  
Senior Investigator Neurogenomics Division, Tgen

Dominique Hoelzinger  
Investigator Neurogenomics, Tgen

Dr. Huan Liu  
Professor, ASU Department of Computer Science

Maulik Shah  
Bioinformatics Programmer, Tgen

## SELECTED BIBLIOGRAPHY

**Agrawal, R., Srikant, R.,** *Fast Algorithms for Mining Association Rules, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, 1994.*

**Bailey, T., Elkan, C.,** *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.*

**Bailey, T., Gribskov, M.,** *Combining evidence using p-values: application to sequence homology searches, Bioinformatics, Vol. 14, pp. 48-54, 1998.*

**Matys, V., Fricke, E., Geffers, R.,** *TRANSFAC®: transcriptional regulation, from patterns to profiles, Nucleic Acids Research, 2003, Vol. 31, No. 1 374-378.*

**Schwartz, S., Kent, J.,** *Human–Mouse Alignments with BLASTZ, Genome Research, Vol 13, Issue 1, 103-107, January 2003.*

**Schwartz, S., Zhang, Z.,** *PipMakerA Web Server for Aligning Two Genomic DNA Sequences, Genome Research, Vol. 10, Issue 4, 577-586, April 2000.*

**Smit, AFA, Hubley, R & Green, P.,** *RepeatMasker Open-3.0. 1996-2004*  
<<http://www.repeatmasker.org>>.

**Zhang, M.Q., Marr, T.G.,** *A weight array method for splicing signal analysis, Computational Applied Bioscience, 1993 Oct;9(5):499-509.*