Computational Investigation of Gene Regulatory Elements

Ryan Weddle
Computational Biosciences
Internship Presentation
12/15/2004
Table of Contents

● Introduction . . . . . 3
● Goals . . . . . 9
● Methods . . . . . 12
● Results . . . . . 21
● Discussion . . . . . 37
● Acknowledgements . . 43
Introduction
Invasive Glioma

- Glioma is a particularly devastating type of brain cancer caused by mutations to glial cells.
- While tumors may be treated through traditional means such as chemo and radiation therapies, these means are less effective at preventing spread and recurrence.
- This is due to the fact that invasive glioma migrates into other parts of the brain by phenotypically different invasive cells.
- These cells are not rapidly dividing and are, thus, less effected by traditional anti-cancer therapy.
Different Tumor Cells

- Tumor composed of core and periphery
- Motile cells are more prevalent in periphery
- Laser capture microdissection used to separate cell populations
What makes them different?

- Microarray analysis performed indicated a set of 15 differentially expressed genes.
- The differential levels of mRNA between the two cell populations were verified with qPCR analysis.
What does this mean?

- When a set of genes are differentially expressed in this manner, it is often hypothesized that they may be co-regulated.
- If they are co-regulated, then understanding their regulation is useful if we wish to prevent their function through some therapeutic means.
Gene Regulation

- Eukaryotic gene regulation is much more complicated than bacterial gene regulation.
- Takes place on several levels:
  - Chromatin remodeling
  - Transcriptional control
  - Message control
  - Translational control
- We are hope to understand the transcriptional control through computational means.
Project Goals
Exploratory Investigation

This project aims to gain understanding of the mechanisms that regulate these differentially expressed genes.

- Leverage sequence data
- Investigate known methods
- Investigate new methods
- Generate and test hypotheses
Leveraging Sequence Data

Two senses in which we are taking advantage of the DNA sequence resources now available:

- Searching genomic sequence data around our genes for transcription factor binding sites
- Using sequence data from multiple genomes to narrow our search
Methods
Investigating Known Methods

- Phylogenetic Footprinting
- Transfac Database
- Pattern Detection Algorithms
- Association Rule Mining
Phylogenetic Footprinting

- Look at sequence which has been conserved over evolutionary time:
  - Ignore coding sequences
  - Ignore known repeating sequences
- Hypothesis is that conserved elements are under selective pressure due to some functional role.
- We used PipMaker to create visualizations, and the blastz software program to compute ungapped alignments.
- Due to limited availability at onset of project, we used only human and mouse genomes.
Example: BCL2L2 Gene Pip

- Black regions are ungapped alignments:
  - Human vs mouse
  - Long segments often codons
  - Notice some upstream conservation
- Percent identity indicated by y-axis.
Transfac Database

- Database of known transcription factor binding sites
  - Catalogues known occurrences
  - Represent TFBS by consensus sequences and weight matrix methods
Matrix Example: TATA

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>3.00</td>
<td>3.89</td>
<td>2.48</td>
<td>3.99</td>
<td>N</td>
</tr>
<tr>
<td>02</td>
<td>0.00</td>
<td>10.19</td>
<td>2.66</td>
<td>0.52</td>
<td>C</td>
</tr>
<tr>
<td>03</td>
<td>0.33</td>
<td>3.33</td>
<td>0.00</td>
<td>9.71</td>
<td>T</td>
</tr>
<tr>
<td>04</td>
<td>9.76</td>
<td>0.00</td>
<td>0.00</td>
<td>3.61</td>
<td>A</td>
</tr>
<tr>
<td>05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13.36</td>
<td>T</td>
</tr>
<tr>
<td>06</td>
<td>13.36</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>A</td>
</tr>
<tr>
<td>07</td>
<td>12.40</td>
<td>0.00</td>
<td>0.00</td>
<td>0.96</td>
<td>A</td>
</tr>
<tr>
<td>08</td>
<td>13.36</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>A</td>
</tr>
<tr>
<td>09</td>
<td>13.36</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>3.92</td>
<td>1.36</td>
<td>6.11</td>
<td>1.97</td>
<td>R</td>
</tr>
</tbody>
</table>

BA  total weight of sequences: 13.36

CC  consind generated matrix (random_expectation: 0.30)
Transfac Utilization

- We can use Transfac to scan DNA sequences:
  - Find potential occurrences
  - Different scores for different quality of matches
- Cannot be used to find novel binding sites, only novel occurrences of known binding sites.
- Useful tool, but too noisy to be relied on in automated processes.
Pattern Detection Algorithms

- Pattern detection algorithms are useful when we are looking for novel motifs.
- We used the MEME/MAST tools to search our conserved sequences for novel motifs:
  - Most interesting result was an already known splice sequence
  - MEME works best when you know how many occurrences you are expecting and where you are expecting them.
Association Rule Mining

- ARM is a mechanism for finding rules about association between different elements.
  - Classical example is “market basket analysis”
  - Here we are interested in any interesting patterns in the occurrence of TFBS identified by Transfac in our conserved sequences.

- Results in many low quality rules:
  - Typically infrequent or low confidence
  - Best rules found due to overlapping putative binding sites - little informational content
Exploratory Investigation Results
Investigating Novel Methods

您同意 existing methods had shortcomings when applied to our dataset:

- Transfac highly uncertain
- Pattern detection and association rule mining failed to yield interesting results
- Too few elements for meaningful clustering, etc.

How can we reframe the problem?
Scaling It All Up

- Association rule mining is intended for large databases.
  - Our gene/TF universe was probably too small to result in interesting rules.
- What if we could scale it up?
  - Look at every subsequence up to a certain length in each genomic region
  - Determine identity between short sequences by allowing slight mismatch
Kmer Analysis

 ARM can be modified to find very low support rules that have high certainty - the “needle in haystack.”

 We can build a database of all TFBS sized short sequences in our conserved sequence data:
  ◦ Mine this database for association rules
  ◦ Interesting rules might indicate functional relationships.
Building the Kmer Database

- Sequence data for each gene was obtained from both mouse and human genomes
  - Repeat sequences and coding regions were masked out.
- Kmer library for all 6-11mers with several degrees of mismatch was constructed
  - 150,000 occurrences of 80,000 unique kmers
  - 550MB on disk
  - 40MB when we exclude all but perfect matches
Refining the Kmer Database

- This is still a very large database!
  - Likely to result in many rules
  - Hard to analyze
  - How can we easily measure the similarity within this database, before devoting time to implementing new algorithms?

- Narrow database to include only exactly matching 11mers
Research Hypothesis

● “There is more short sequence similarity, as measured by exactly matching 11mers, in our target sequence corpus, than would be expected from random sequence data.”

● If we can confirm this hypothesis, we can assert that there is interesting informational content at the sequence level.
  ○ Worthwhile to investigate further
Randomized Sequence Data

- We needed a basis for comparison to determine whether the short sequence similarity observed in our data set was significant.
  - Generate random sequence data that maintains the same nucleotide bias for each sequence fragment
  - Perform kmer analyses on each of these random trials
  - 100 trials in total
Research Hypothesis Results

Randomly Generated Sequences
Summary Statistics

☀ 11mer distributions calculated for both
  ☀ Uniform nucleotide distribution
  ☀ Same distribution as in target data
    ☰ A=26.7% C=23.0% G=26.1% T=24.2%

☀ Z-test:
  ☀ Is our observed count of 73 11mers higher than the population mean?
    ☰ Z score = -46
    ☰ P-value < 10^-6

<table>
<thead>
<tr>
<th>11mers</th>
<th>Uniform</th>
<th>Controlled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>33.65</td>
<td>36.28</td>
</tr>
<tr>
<td>Median</td>
<td>33.50</td>
<td>36.00</td>
</tr>
<tr>
<td>Variance</td>
<td>57.89</td>
<td>62.51</td>
</tr>
<tr>
<td>StDev</td>
<td>7.61</td>
<td>7.91</td>
</tr>
<tr>
<td>Min</td>
<td>16.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Max</td>
<td>49.00</td>
<td>58.00</td>
</tr>
</tbody>
</table>
Hypothesis Revisited

- Results looked promising..
- However, they depended on assumptions about random sequence data.
- Therefore, we revised our hypothesis:
  - “There is more short sequence similarity, as measured by exactly matching 11mers, in our target sequence corpus, than would be found by randomly sampling sets of genes from the human and mouse genomes.”
- Confirming this hypothesis would provide concrete evidence that our observed 11mer similarity constituted a meaningful departure from the norm.
Analyzing Random Genes

- Downloaded all human-mouse homologs from EMBL
- Performed pre-processing on all homolog pairs
  - Repeatmasking
  - Blastz for phylogenetic footprinting
- Randomly selected 100 sets of genes
- Performed 11mer analysis on every set
- Catalogued results
Research Hypothesis Results

Randomly Selected Sequences
Distributions Overlay

73 observed 11mers
Comparing the Distributions

- All distributions appear normal.
- 73 observed 11mer matches are clearly
  - More occurrences than expected from random sequence
  - Much fewer than expected from randomly selected genes
- What’s going on here?
Discussion
Conclusions

- 73 observed 11mer matches are anecdotally interesting.
  - Transfac matches for TATA, various TFs
- Our most exhaustive results indicate that, however, we cannot claim that the number of matches are statistically significant.
- But, there are more variables involved in the final analysis, which could be controlled for in further analyses.
Possible Confounding Factors

- Amount of conserved sequence may differ due to:
  - Percent conservation
  - Size of genes
- Controlled for in random sequence generation, but not in random gene selection
- Assumes all genes are comparable
- Controlling for these factors could be a good avenue for future research
Questioning Assumptions

- Everything rests on the assumption that our target set of genes is co-regulated by common elements at the DNA sequence level.
  - Further assumption that regulatory mechanism is local to the genes
  - What about chromatin and its role in regulation?
Suggestions for Future Work

- It would be useful to repeat the final tests while controlling for gene size and conservation.
- Consider testing these same methods on an already well characterized set of co-regulated genes, rather than on an investigative data set.
- Research methods for taking chromatin and DNA sequence structure into account.
Things I Learned

- In exploratory investigations, Perl is your best friend.
  - Much of this would have been impossible to do manually.
  - Perl really is faster for rapid prototyping when you don’t know in advanced what your needs will be.
- You can try new methods on old data, or old methods on new data, but developing new methods on new data is difficult.
Acknowledgements

- Dr. Jeff Touchman . Tgen, ASU
- Dr. Phillip Stafford . Tgen, ASU
- Dr. Rosemary Renaut . ASU
- Dr. Michael Berens . Tgen
- Dr. Huan Liu . ASU
- Dominique Hoelzinger . Tgen
- Maulik Shah . Tgen, ASU