

Research Summary

Pouya Kheradpour

February 9, 2005

<https://netfiles.uiuc.edu/kheradpo/www/beespace-pres.pdf>

cDNA Microarrays

While each cell in an organism has the same genome, the amount of mRNA for each gene differs and determines what proteins are being synthesised by a cell.

Microarray technology allows for measuring the relative abundance of mRNA for several thousand genes simultaneously using light intensity measurements.

Ultimately, we want the expression level for each gene in each sample. Statistical techniques and various normalizations must be employed to obtain this.

Genome Sequence

The genome of an organism refers to its genetic blueprint comprised only of the letters A, T, G and C.

Genome sequencing today requires the computational assembly of millions of short shotgun sequences.

Theoretically, the shotgun sequences are equally likely to come from any portion of the genome. Unfortunately, in practice, this is not always the case.

The previous draft of the honey bee genome had serious problems. New draft (6x coverage) is much improved, but still not on the same level as some other genomes such as human, mouse or fruit fly.

After assembly, it will take a couple months for the genes in the honey bee genome to be computationally annotated.

Research Focus

Bioinformatics has become increasingly important as massive amounts of sequence and expression data have been introduced requiring computational analysis.

The focus of our research has been to develop algorithmic approaches to finding features that contribute to the differential expression of genes.

We have been using both microarray expression and genome sequence data to find these features. Typically, we try our methods first on yeast or fruit fly where more data is often available or the organism is simpler and then move on to the honey bee.

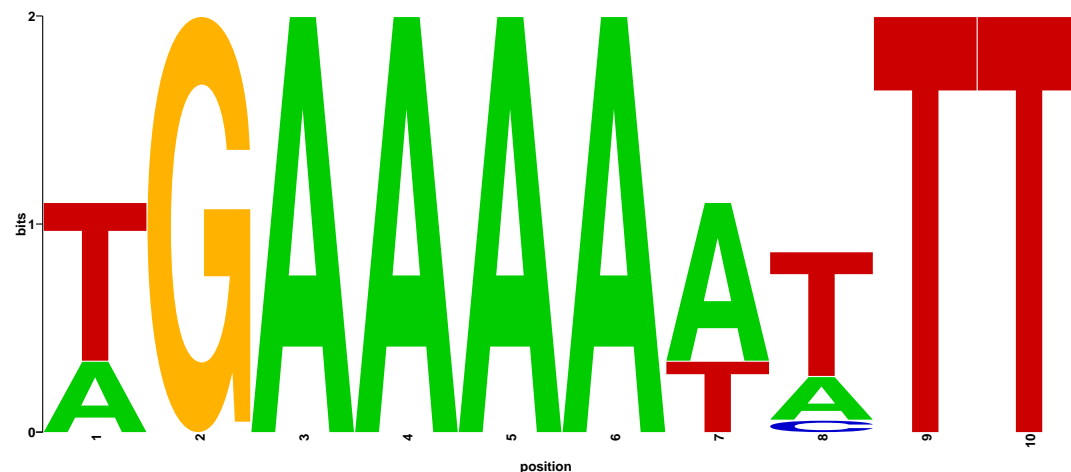
Transcription Factor Binding Sites

Found in the 5' upstream region and perhaps in the first intron of a gene.

How far away from the start of the gene these can be found is not well understood and differs depending on the organism.

Further, the binding motif can be very short (5 bases) and are often degenerate.

These properties can make identifying these binding sites very difficult.



Related Work

Michael A. Beer and Saeed Tavazoie. "Predicting gene expression from sequence." *Cell*, 2004, 117: 185-198.

Attempts to computationally discover network driving expression of yeast genes.

Method involves k-means clustering genes using microarray expression data followed by training a bayesian network to identify the appropriate cluster using motifs found in their upstream region.

Unfortunately, cross validation was not strictly observed during motif finding, resulting in misleadingly high performance. Also, typically the bayesian networks discovered were very simple.

However, there are many opportunities for variations and perhaps improvements. Clustering can be done using several algorithms and support vector machines can replace the custom bayesian network.

P-Value

Rather than develop complex bayesian network heuristics a p-value approach may be appropriate and we found much simpler.

A positive and negative set of genes must be supplied. These sets can come from, for example, clustering, ontological knowledge or from a PCA approach (more on this later).

Positive set of genes are randomly broken into training and test sets.

Motif finding (using AlignACE or other algorithm) is applied to training positive set.

Test positive and negative sets are scanned for the identified motifs.

All motifs that are found in statistically different proportions are flagged.

Flagged motifs can be compared to known motifs to evaluate the performance of the method.

P-Value, Continued

Typically, this procedure is run tens to hundreds of times to obtain different positive sets and AlignACE motifs.

Fortunately, this is feasible because each run can take as little as 5 minutes when small training sets are used.

Perhaps pairwise combinations (with AND or OR logic) of found motifs can be used.

Method appears to do well on simple model of yeast. Further analysis is needed to evaluate performance on honey bee.

Principle Component Clustering

One gene may contain numerous enhancer motifs. However, many clustering algorithms, such as k-means, only allow a gene to be in one cluster.

To perform principle component clustering first we transform sample space into PCA space. PCA space is the projection such that each dimension contains the maximal variance such that it is orthogonal to all previous dimensions.

For each principle component we take the top and bottom scoring genes to each be a positive set. We use the genes with a middle (near 0) score to be the corresponding negative gene set.

Neighborhood Effect

Genes close to one another in the genome appear to have similar expression levels. While some is already known about this effect, it doesn't appear to have been fully quantified.

In the fruit fly we have found about 0.11 average correlation between the expression pattern of genes next to each other in a chromosome whereas the average correlation of a random pair of genes appears to be approximately 0.

Further analysis is required for the honey bee, but will be possible once annotation of the new assembly is completed.

Linear Effects Model

Idea: Can the effect of transcription factors be modeled as a linear system?

Identify motifs and which genes they are found in (using, e.g., AlignACE).

Find samples in which a motif appears to be “active” by seeing if genes with a motif are heavily expressed in a sample.

Use linear least squares to determine the linear effect of each motif.

Least squares system appears to be poorly conditioned. Can better motifs, a nonlinear system, etc. provide better results?