BIOINFORMATICS
Vol. 00 no. 00 2005
Pages 1–3

Runs Tests to Detect Co-Regulation Patterns in Wood Tissues
Zhen Li,* Yao Li and George Casella
Department of Statistics, University of Florida, Gainesville, FL 32611, United States

ABSTRACT

Motivation: Whole genome microarrays provide gene expression data taken from different wood tissues. An efficient and easily interpretable method is needed to assess co-regulation of adjacent genes.

Results: The statistical analysis methods proposed here successfully analyzed the co-regulation of genes in several tissue types and detected different gene expression patterns. These methods can be applied to analyze the gene expressions in other eukaryotes.

Availability: Program is available from http://plaza.ufl.edu/zhenli79/

Contact: zhenli79@ufl.edu

1 INTRODUCTION

It is well known that differential transcriptional regulation determines different gene expression among tissue types. Previous work of Spellman et al., 2002 showed that adjacent genes have similar expression patterns in Drosophila. Here we want to assess the changes in gene expression in different wood tissues, as a result of tissue specification or treatments such as wounding, hormone applications or nitrogen fertilization. Our data has gene expression data taken from different wood tissues. An efficient and easily interpretable method is needed to assess co-regulation of adjacent genes.

2 APPROACH

We take a three part approach to identifying sets of correlated genes, with each part becoming more focused on particular subsets of genes. First, we adapt a procedure of O’Brien et al., 1985 to assess whether the overall gene expression pattern differs from that of randomly placed genes. This allows either a linkage-group wide or genome-wide control of errors. If significance is found, then we measure the pattern of the runs (sequences of co-regulated genes) to see how this pattern deviates from randomness. If there is co-regulation, we would expect the observed run lengths to be upwardly biased. Third, we use a sliding window to examine all adjacent sets of genes and measure the deviation from randomness in each window. This last step allows us to pinpoint the regions in the genome where there is co-regulation.

3 METHODS

To perform the overall test of co-regulation in each linkage group, each gene is identified with 0 if its expression level is nonsignificant, and 1 if it is. We then search through the sequence of 0s and 1s to see if there are patterns, or runs. Thus, the null hypothesis is that the pattern of 0s and 1s is random. To detect deviation from random expression, we use a runs test first proposed by O’Brien et al., 1985. The test is based on run lengths, where a run is defined to be a succession of the same digit, bordered by different digits, and the length of a run is the number of digits in that run. For example, consider a sequence of 1-1-0-1-1-0-1-0-0-0-1-0-1-1. Here there are 7 runs with run lengths 2, 1, 3, 3, 1, 1, 2. The 1st, 3rd, 5th, 7th runs are composed of 1s and the others are composed of 0s. The runs test requires the following notation. For expressed genes:

\[
\begin{align*}
    n_1 &= \text{total number of 1s} \\
    r_1 &= \text{total number of runs composed of 1s} \\
    s_1^2 &= \text{sample variance of the runs of 1s} \\
    c_1 &= \frac{(r_1 + 1)(r_1 + 2)(r_1 + 3)}{2r_1(n_1 - r_1 - 1)(n_1 + 1)}
\end{align*}
\]

The notation for unexpressed genes, which is given the subscript 0, is defined in similar fashion. The test statistic is:

\[
\chi^2 = c_1 s_1^2 + c_2 s_2^2.
\]

O’Brien et al., 1985 derived the asymptotic distribution of \(c_1^2\) of (2) under the null hypothesis that the sequence of genes has no co-regulations and thus the genes are randomly expressed or randomly unexpressed. However, here we will use a permutation - bootstrap approach to obtain the null distribution, and we generate 25000 independent data sets under the null hypothesis, where each data set is a sequence of independent Bernoulli trials. The length of the sequence is the same as that of the observed sequence and the proportion \(p\) of 1s in the null sequence of Bernoulli trials is equal to the proportion of 1s in the observed data, since under null hypothesis, this proportion does not change. We then calculate the \(\chi^2\) statistic of (2) for each generated data set and compare those values to the \(\chi^2\) value of the observed data, obtaining a p-value for our test of no co-regulation.

To assess the pattern of runs, we again generate 25000 data sets under the null hypothesis, and for each data set we calculate the number of runs of length 1, the number of runs of length 2, up to the longest run in the data set. This gives us a null distribution for each run length. Then from the observed data, we calculate the number of runs of length 1, the number of runs of length 2, etc. For each run length we can now assess whether the observed value of the run length deviates from what is to be expected under the null hypothesis that gene expressions are not co-regulated.

Finally, for each run length, if the null hypothesis is rejected, we can find the position of the runs of that length by using a sliding window to detect the exact position of the non-randomly expressed genes in the sequence. We experimented with different window lengths, and found that windows less than 100 did not result in adequate power. For each window we generate 25000 data sets under the null hypothesis, that is, we generate 100 Bernoulli trials with success probability equal to the proportion of 1s in the entire linkage group. For every segment of 100 genes, if there were no co-expressed genes, we expect the same proportion of expressed genes as the entire linkage group. We then calculate the \(\chi^2\) statistic (2) for each sliding window and compare it to the null distribution of the 100 gene segment. This last

* to whom correspondence should be addressed

inference is adjusted by using the false discovery rate of Benjamini et al., 1995.

We can also modify these methods to deal with genes from more than one wood tissue type, by constructing a sequence of pair-wise comparisons. For example, suppose the sequence of gene expressions in Young Leaf Tissue is 1-0-1-1-0-0-0-0-0-0 and the sequence of expressions in Internode Tissue is 0-0-1-1-0-1-0-0. We calculate the element-wise absolute difference between these tissue types to obtain the sequence 1-0-0-0-0-0-1-0, where the 1 denotes differential gene expression, and the 0 denotes similar gene expression. In this example, with a long run of 0s, we intuitively conclude that these two sequences have similar expression patterns. We can perform the three tests described above on such sequences of absolute difference for the real-data tissue expressions.

4 DISCUSSION

Our poplar data set is composed of 5 wood tissue types and 19 linkage groups in each tissue type. A gene expression value is denoted by the transition matrix

$$P = \begin{pmatrix} 1 - p & p \\ q & 1 - q \end{pmatrix} \quad (3)$$

If we define $$p^* = \frac{p}{p+q}$$ and $$q^* = p + q$$, then the stationary distribution has $$P(X_{t+1} = 1) = p^*$$, $$\text{Corr}(X_t, X_{t+1}) = 1 - q^*$$, and the null hypothesis has $$p = 1 - q$$, which gives zero correlation.

We further investigated the length of the runs, and found out that most of the co-regulated linkage groups have similar patterns: short runs tend to appear fewer times than was expected under randomness, while long runs tend to appear more than expected. Figure 1 shows the results of the pattern of run lengths test for linkage group 13. We see that the shorter runs are below the expected run length line, while the longer runs are above it.

To locate the areas of co-regulation, we used the above mentioned sliding window test. However, we first did a power study to determine an appropriate window length to give us adequate power. Figure 2 shows the results of the powers for different window lengths. The pattern of power in $$p^*$$ has a minimum at 1/2, and increases symmetrically. As the correlation increases, the power also increases. We found that a sliding window length of 100 gave acceptable power, as shown in Table 2.

As an example of the power evaluations we performed, one simulation used the following values:

1. For $$n = 25, 50, 100$$
2. For $$p^* = 0.1, 0.25, 0.5, 0.75$$
   a. Set $$q^* = 1$$ and generate 5000 samples to get the null distribution of $$\chi^2$$.
   b. For $$q^* = 0.25, 0.5, 0.75$$ generate 5000 samples and get the proportion of times $$\chi^2$$ is over the .01 cutoff, which is the power.

The pattern of power in $$p^*$$ has minimum at 1/2, and increase symmetrically. As the correlation increases, the power also increases. We found that a sliding window length of 100 gave acceptable power, as shown in Table 2.

An example of the sliding window test output is given in Figure 2. The small $$p$$-values identify the start of window containing genes that display significant co-regulation. The overall window is calibrated with the FDR $$Q$$-value to maintain the error rate.

5 CONCLUSION

1. The overall test of randomness shows the existence of regions of co-regulated genes on linkage groups, and gives us direction to further pinpoint the position of the genes.
2. For our data, runs length test shows that in the co-regulated linkage groups, there tend to be fewer numbers of runs than expected for short lengths and more runs than what is expected for long length.
3. The power of the sliding window test for this data set is not very high, but for window length 100 we still were able to find areas where there is significant co-regulation.

ACKNOWLEDGEMENT

We thank John Davis and Matias Kirst of the School of Forestry for providing us with the data.

REFERENCES

