

# A Two-dimensional Regression Tree Approach to the Modeling of Gene Expression Regulations

Jianhua Ruan <sup>1</sup> and Weixiong Zhang <sup>2</sup>

**Keywords:** gene regulation, microarray analysis, data mining, multivariate regression tree

## 1 Introduction.

Gene transcription depends on the binding of some transcription factors (TFs) to promoters, and the expressions of TFs. Traditional approaches to studying transcriptional regulations usually model these dependencies separately, i.e., either from promoters to gene expressions (e.g., [4]), or from the expressions of TFs to the expressions of genes (e.g., [6]).

Here we propose a new method, called Two-Dimensional Regression Tree (BDTree), to model gene expressions from both promoter sequences and TF expressions. The model built by BDTree provides testable hypotheses about condition-specific binding motifs and TFs for genes, and can also be used to predict the expression levels of unknown genes under unknown conditions, given some attributes of the genes and conditions.

## 2 Algorithms.

Suppose that we are given the expression matrix for  $m$  genes under  $n$  conditions,  $\mathbf{E} = (e_{ij})_{m \times n}$ , where  $e_{ij}$  is the expression level of the  $i$ th gene under the  $j$ th condition. We are also given the scores of  $p$  candidate motifs on the promoter of each gene,  $s_{i1}, s_{i2}, \dots, s_{ip}, i \in [1..m]$ , and the expression levels of  $q$  putative TFs under these conditions,  $t_{1j}, t_{2j}, \dots, t_{qj}, j \in [1..n]$ . We call each motif a *gene attribute*, and each TF a *condition attribute*. The BDTree algorithm recursively partitions  $\mathbf{E}$  horizontally or vertically. An eligible horizontal partition splits genes into two subsets,  $I$  and  $J$ , such that  $s_{ik} < s_{jk}$ , for all  $i \in I, j \in J$  and some  $k \in [1..p]$ . Similarly, an eligible vertical partition splits conditions into two subsets,  $I$  and  $J$ , such that  $t_{ki} < t_{kj}$ , for all  $i \in I, j \in J$  and some  $k \in [1..q]$ . The outline of the algorithm is as follows:

1. Initially there is only the root node containing all genes and conditions.
2. If the stopping criterion has not yet been met, examine every eligible horizontal or vertical partition and measure the goodness of the partition.
3. Choose the best partition and create two child nodes for the current node.
4. Repeat steps 2 and 3 for each child node.
5. Post-prune and cross-validate the tree.

The goodness of a partition is defined to reflect the within-node homogeneity of expression submatrices, where homogeneity is measured by a sum-of-squares function as in [1]. This measurement does not always look for constant-valued submatrices; a matrix is homogeneous if the variance within it can be explained by the variance of its rows and columns. The algorithm effectively divides  $\mathbf{E}$  into many small homogeneous blocks, which is analogous to Biclustering [1]. However, the partitioning of genes and conditions in BDTree is supervised by some intrinsic attributes of the genes and conditions, i.e., motifs and TFs. In contrast, the partitioning in Biclustering is unconstrained by these attributes.

---

<sup>1</sup>Department of Computer Science and Engineering, Washington University in St Louis, St Louis, Missouri, USA. E-mail: [jruan@cse.wustl.edu](mailto:jruan@cse.wustl.edu)

<sup>2</sup>Department of Computer Science and Engineering and Department of Genetics, Washington University in St Louis, St Louis, Missouri, USA. E-mail: [zhang@cse.wustl.edu](mailto:zhang@cse.wustl.edu)

### 3 Results.

We first applied BDTree to the microarray data of 800 yeast cell cycle genes under 77 time points [7]. We combined ChIP-chip data [3], all pentamers, and 356 known and putative motifs [5] as gene attributes, and used 475 putative TFs as condition attributes [6]. A portion of the resulting tree is shown in Figure 1. To identify important attributes, we ranked all attributes by their significance scores [4]. Surprisingly, 18 of the top 20 gene attributes and 15 of the top 20 condition attributes are well known for their roles in regulating cell cycle genes, including eight of the nine essential cell cycle TFs. We also applied BDTree to the expression of  $\sim 4000$  yeast genes under 173 stress conditions [2], and found many known motifs and TFs for stress responses.

We estimated the prediction accuracy of BDTree with 10-fold cross validations and compared it to a  $K$ -nearest-neighbors method. The correlation coefficient between the actual expression levels and the values predicted by BDTree is 0.31, which is significantly better than 0.18 for the  $K$ -nearest-neighbors method ( $p < 10^{-20}$ ).

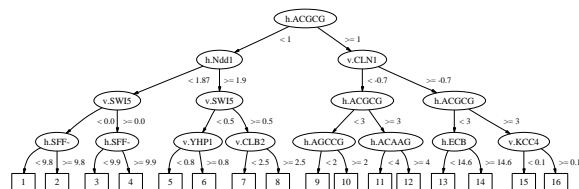


Figure 1: Regression tree learned for 800 yeast cell cycle genes. The tree was pruned to show only the top portion. Each circle represents an attribute. Prefix ‘h’ and ‘v’ represent horizontal and vertical splits, respectively. The labels associated with branches are attribute threshold values for partitioning the training data. Each square represents a submatrix of the expression matrix.

### References

- [1] Cheng, Y. and Church, G.M. 2000. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol.* 8:93–103.
- [2] Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. and Brown, P.O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11:4241–4257.
- [3] Lee, T.L., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. 2000. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science* 298:799–804.
- [4] Phuong, T.M., Lee, D. and Lee, K.H. 2004. Regression trees for regulatory element identification. *Bioinformatics* 20:750–757.
- [5] Pilpel, Y., Sudarsanam, P. and Church, G.M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29:153–159.
- [6] Segal, E., Shapira, M., Regev, A., Pe’er, D., Botstein, D., Koller, D. and Friedman, N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34:166–176.
- [7] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273–3297.