BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 5 2008, pages 719–720
doi:10.1093/bioinformatics/btm563

Gene expression

Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R

Peter Langfelder[1,†], Bin Zhang[2,†] and Steve Horvath[1,*]

[1]Department of Human Genetics, University of California at Los Angeles, CA 90095-7088 and
[2]Rosetta Inpharmatics-Merck Research Laboratories, Seattle, WA, USA

Received and revised on September 12, 2007; accepted on November 6, 2007

Advance Access publication November 16, 2007

Associate Editor: Trey Ideker

ABSTRACT

Summary: Hierarchical clustering is a widely used method for detecting clusters in genomic data. Clusters are defined by cutting branches off the dendrogram. A common but inflexible method uses a constant height cutoff value; this method exhibits suboptimal performance on complicated dendrograms. We present the Dynamic Tree Cut R package that implements novel dynamic branch cutting methods for detecting clusters in a dendrogram depending on their shape. Compared to the constant height cutoff method, our techniques offer the following advantages: (1) they are capable of identifying nested clusters; (2) they are flexible—cluster shape parameters can be tuned to suit the application at hand; (3) they are suitable for automation; and (4) they can optionally combine the advantages of hierarchical clustering and partitioning around medoids, giving better detection of outliers. We illustrate the use of these methods by applying them to protein–protein interaction network data and to a simulated gene expression data set.

Availability: The Dynamic Tree Cut method is implemented in an R package available at http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/BranchCutting

Contact: stevitihit@yahoo.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Detecting groups (clusters) of closely related objects is an important problem in bioinformatics and data mining in general. Many clustering methods exist in the literature (Hastic *et al*., 2001; Kaufman and Rousseeuw, 1990). We focus on hierarchical clustering, but our methods are useful for any clustering procedure that results in a dendrogram (cluster tree). Hierarchical clustering organizes objects into a dendrogram whose branches are the desired clusters. The process of cluster detection is referred to as tree cutting, branch cutting, or branch pruning. The most common tree cut method, which we refer to as the 'static' tree cut, defines each contiguous branch below a fixed height cutoff a separate cluster. The structure of cluster joining heights often poses a challenge to cluster definition. While distinct clusters may be recognizable by visual inspection, computational cluster definition by a static cut does not always identify clusters correctly. To address this challenge, we have developed a novel tree cut method based on analyzing the shape of the branches of a dendrogram. As a motivating example, consider Figure 1A that shows a dendrogram for cluster detection in a protein–protein interaction network in *Drosophila*. The Dynamic Tree Cut method succeeds at identifying branches that could not have been identified using the static cut method. The found clusters are highly significantly enriched with known gene ontologies (Dong and Horvath, 2007) which provides indirect evidence that the resulting clusters are biologically meaningful.

2 ALGORITHM AND IMPLEMENTATION

We provide only a brief summary of the Dynamic Tree Cut method here; a detailed description is given in the Supplementary Material. To provide more flexibility, we present two variants of the method. The first variant, called the 'Dynamic Tree' cut, is a top-down algorithm that relies solely on the dendrogram. This variant has been used to identify biologically meaningful gene clusters in microarray data from several species such as yeast (Carlson *et al*., 2006; Dong and Horvath, 2007) and mouse (Ghazalpour *et al*., 2006), but has not previously been systematically described nor made publicly available. The algorithm implements an adaptive, iterative process of cluster decomposition and combination and stops when the number of clusters becomes stable. It starts by obtaining a few large clusters by the static tree cut. The joining heights of each cluster are analyzed for a characteristic pattern of fluctuations (see Supplementary Material for details) indicating a sub-cluster structure; clusters exhibiting this pattern are recursively split. To avoid over-splitting, very small clusters are joined to their neighboring major clusters.

The second variant, called the 'Dynamic Hybrid' cut, is a bottom-up algorithm that improves the detection of outlying members of each cluster. The detection proceeds in two steps. First, the method identifies preliminary clusters as branches that satisfy the following criteria: (1) they contain a certain minimum number of objects; (2) objects too far from a cluster

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

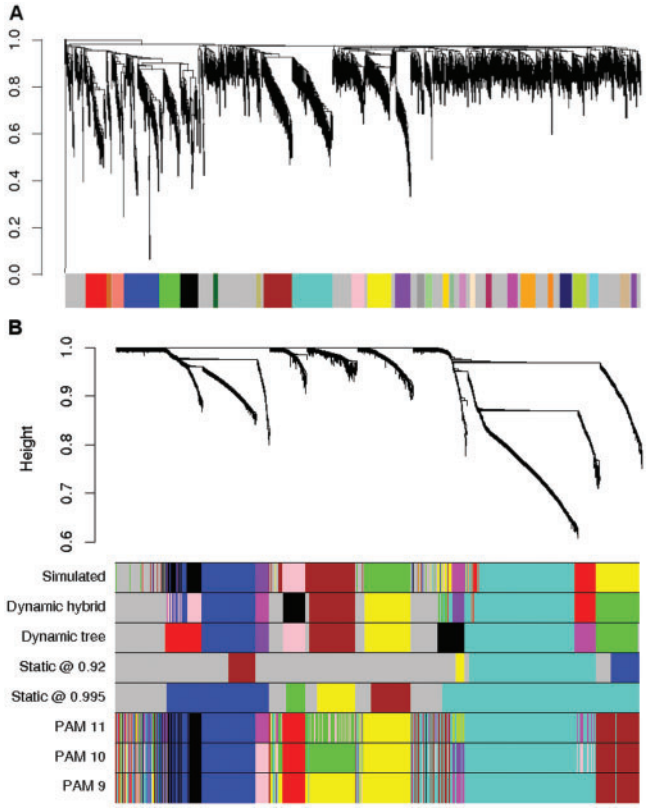

Fig. 1. (**A**) Average linkage hierarchical clustering using the Topological Overlap Matrix (Yip and Horvath, 2007) and the Dynamic Tree cut applied to the protein–protein interaction network of *Drosophila* (PPI data from BioGRID, www.thebiogrid.org). Module assignment is depicted by the row of color immediately below the dendrogram, with gray representing unassigned proteins. A functional enrichment analysis has shown that the clusters are significantly enriched with known gene ontologies (Dong and Horvath, 2007). Note that a fixed height cutoff would not be able to identify many of the shown clusters. (**B**) Hierarchical cluster tree and various cluster detection methods applied to a simulated gene expression data set. The color bands below the dendrogram show the cluster membership according to different clustering methods. The color gray is reserved for genes outside any proper cluster, i.e., the tree cut methods allow for unassigned objects. The first color band 'Simulated' shows the simulated true cluster membership; color bands 'Dynamic Hybrid' and 'Dynamic Tree' show the results of the proposed tree cutting methods; the color band 'Static @ 0.92' shows the results of the standard, constant height cut-off method at height 0.92. The height refers to the y axis of the dendrogram. The color band 'PAM 11' shows the results of $k = 11$ medoid clustering.

are excluded from it even if they belong to the same branch of the dendrogram; (3) each cluster must be distinct from its surroundings and (4) the core of each cluster, defined as the tip of the branch, should be tightly connected. In the second step, all previously unassigned objects are tested for sufficient proximity to preliminary clusters; if the nearest cluster is close enough, the object is assigned to that cluster, see the Supplementary Material. Since Partitioning Around Medoids (PAM; kaufman and Rousseeuw, 1990) also involves assigning objects to their closest medoids, the Dynamic Hybrid variant can be considered a hybrid of hierarchical clustering and modified PAM.

3 DISCUSSION

Many clustering procedures have been developed for the analysis of microarray data (Dembele and Kastner, 2003; Ghosh and Chinnaiyan, 2002; Thalamuthu *et al*., 2006; van der Laan and Pollard, 2003). Our method could be useful for any procedure that results in a dendrogram. In Figure 1B, we report a simulated gene expression data set (detailed description can be found in the Supplementary Material). We simulated 10 nested gene clusters labeled by different colors, shown in the first color band underneath the dendrogram. The results of the methods presented here are shown in correspondingly labeled color bands. Visual inspection shows the Dynamic Hybrid method outperforms the Static height cutoff method whose clusters either contain too few genes or are too coarse. PAM forces all genes into clusters and favors membership in large clusters at the expense of small ones. A quantitative analysis of this example is presented in the Supplementary Material.

The height and shape parameters of the Dynamic Tree Cut method provide improved flexibility for tree cutting. It remains an open research question how to choose optimal cutting parameters or how to estimate the number of clusters in the data set (Dudoit and Fridlyand, 2002). While our default parameter values have worked well in several applications, we recommend to carry out a cluster stability/robustness analysis in practice.

ACKNOWLEDGEMENTS

We thank Ai Li, Jun Dong and Tova Fuller for discussions and suggestions. We acknowledge the grant support from 1U19AI063603-01 and NINDS/NIMH 1U24NS043562-01.

Conflict of Interest: none declared.

REFERENCES

Carlson,M.R. *et al*. (2006) Gene connectivity, function, and sequence conservation: predictions from modularv yeast co-expression networks. *BMC Genomics*, **7**.

Dembele,D. and Kastner,P. (2003) Fuzzy C-means method for clustering microarray data. *Bioinformatics*, **19**, 973–980.

Dong,J. and Horvath,S. (2007). Understanding network concepts in modules. *BMC Syst. Biol*., **1**, 24.

Dudoit,S. and Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol*., **3**, research0036.1–research0036.21.

Ghazalpour,A. *et al*. (2006) Integrating genetics and network analysis to characterize genes related to mouse weight. *PLoS Genet*., **2**, e130.

Ghosh,D. and Chinnaiyan,A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275–286.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning*. Springer, New York.

Kaufman,L. and Rousseeuw,P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., New York.

Thalamuthu,A. *et al*. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

van der Laan,M.J. and Pollard,K.S. (2003) A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *J. Stat. Plan. Inference*, **117**, 275–303.

Yip,A. and Horvath,S. (2007) Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, **8**, 22.