Eurographics Conference on Visualization (EuroVis) (2016)
T. Isenberg and F. Sadlo (Editors)

Hierarchical Clustering with Multiple-Height Branch-Cut Applied to Short Time-Series Gene Expression Data

T. Vogogias[1], J. Kennedy[1] and D. Archambault[2]

[1]Edinburgh Napier University, United Kingdom
[2]Swansea University, United Kingdom

Abstract

Rigid adherence to pre-specified thresholds and static graphical representations can lead to incorrect decisions on merging of clusters. As an alternative to existing automated or semi-automated methods, we developed a visual analytics approach for performing hierarchical clustering analysis of short time-series gene expression data. Dynamic sliders control parameters such as the similarity threshold at which clusters are merged and the level of relative intra-cluster distinctiveness, which can be used to identify "weak-edges" within clusters. An expert user can drill down to further explore the dendrogram and detect nested clusters and outliers. This is done by using the sliders and by pointing and clicking on the representation to cut the branches of the tree in multiple-heights. A prototype of this tool has been developed in collaboration with a small group of biologists for analysing their own datasets. Initial feedback on the tool has been positive.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Viewing algorithms—H.3.3 [Information Search and Retrieval]: Clustering—Information filtering

1. Introduction

Genes that are involved in the same biological process usually follow a similar expression pattern. Therefore, a common technique for reducing complexity in a dataset is to study groups of co-expressed genes, rather than single genes [RAHZ13]. Hierarchical clustering algorithms (HCAs) are often used for classifying genes into separate groups, based on similarities in gene expression levels. This is an unsupervised approach for inferring structure in the data. In the case of time-series gene expression data, each of the clusters corresponds to a distinct temporal profile, or pattern [WWLC08].

The challenge is to identify groups of genes in the hierarchical structure produced by the HCA, which is known as the *dendrogram*. However, every clustering scenario is merely a hypothesis to be tested [EC02]. Automated approaches, such as the *Dynamic Tree Cut (DTC)* [LZH08], use heuristic criteria, which are not unique and produce different clustering results. Semi-automated approaches, on the other hand, integrate prior knowledge into the algorithm [DCMK07, NWN*09]. Hence, the solution is based on assumptions about the data in hand. In the real world, there is no *"one-size-fits-all"* solution and it is common to ignore special characteristics of clusters [KK99]. Within the same dataset some clusters may be dense (high similarity), while some others may be sparse (low similarity). For instance, biologically associated genes may follow a similar expression pattern during the whole experiment, or only for a time period [MABK11, CCKK12]. Therefore, the *'human in the loop'* is needed to visually explore the dendrogram and select potential subsets manually [SM11]. For datasets which consist of homogeneous subsets, deciding a single similarity threshold, which cuts the tree in a unified height, would be sufficient. However, for larger dendrograms, which often consist of heterogeneous subsets, a more effective approach would be to choose multiple similarity thresholds. These thresholds could be applied iteratively until a satisfactory partitioning of the dendrogram is achieved. In other words, the user task could be transformed into finding *where to cut the branches that form different clusters*, by suggesting different clustering scenarios (Figure 1). Similar approaches have been investigated in the past for exploring graph structures, as in [AMA08, AVHK06].

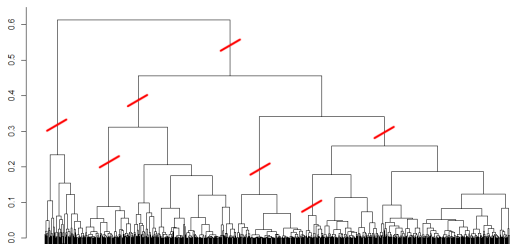

Figure 1: *Multiple-height branch-cut.*

We have developed an interactive tool that enables the user to manually select clusters by applying multiple-height branch-cuts

on demand. There are two types of similarity thresholds: *global* that apply to the whole dendrogram and *local* that only apply to parts of the tree, such as selected branches of interest, enabling a more finely-grained exploration. This is a synergistic approach that combines the strengths of HCAs with the ability of humans to visually detect patterns and anomalies in the data.

2. Previous Work

There are many visualisation tools for exploring patterns in temporal data. However, most of them are either too generic to be applicable to short time-series gene expression data, or only appropriate to serve the analysis needs of particular datasets. For instance, *TimeSearcher* [HS01] is a generic tool for interactively exploring time-series. However, the users need to know in advance the time patterns they are interested in. *Hierarchical Clustering Explorer (HCE)* [SS02] is a tool designed for multivariate gene expression data. However, it only supports a single-height cut-off similarity threshold for performing clustering analysis. Similarly, a technique presented in [CMP09] uses a single-height cut-off value to provide improved visibility by simplifying the dendrogram representation.

In addition, many visualisation approaches focus on exploring phylogenetic trees [HBW*14], which look similar to dendrograms. *Treejuxtaposer* [MGT*03] compares two phylogenetic trees to merge similar parts. Tree comparison techniques are not useful for exploring different merging scenarios in a single dendrogram. To our knowledge, the most relevant visualisation tool for our task is *Pathline* [MWS*10]. This is an interactive tool, developed as a case study for exploring gene expression temporal profiles across different species. It requires synchronised experimental data from three different sources, which unfortunately is rare and expensive to reproduce [CZWT15]. Our approach can be applied to any short time-series, which according to [EBJ06] constitute the 80% of Microarray time-course gene expression datasets.

3. Visual design and UI controls

In the typical top-down dendrogram representation (Figure 1), each edge length is proportional to the overall distinctiveness score of the branch it connects with the rest of the tree [CRV08]. However, for large dendrograms, this convention does not always produce intelligible results, because the hierarchical structure is too large to browse. In our approach we adopted a space efficient radial layout [MR10] and we quantify and control the property of distinctiveness using a dynamic slider [AS94].

The user interface (UI) is composed of two linked view components. The top view constitutes a radial representation of the dendrogram, while the bottom view is a representation of the original short time-series gene expression data, using parallel coordinates [ID91]. A capture of the user interface is shown in Figure 2.

A global similarity threshold can be applied using a dynamic slider. Branches that belong to the same cluster get the same colour. This feature is useful for identifying the main clusters and also for testing the different scenarios that the single-height approach can investigate. A second dynamic slider can be used for identifying sub-clusters, which appear considerably more homogeneous than

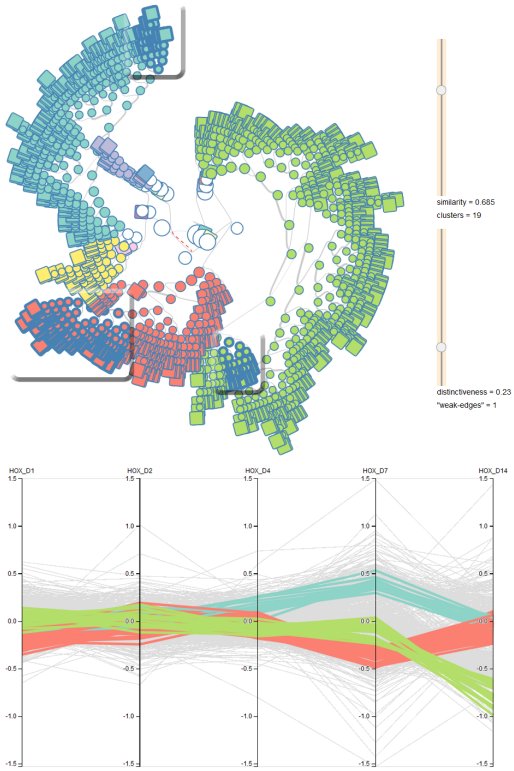

Figure 2: *Three subsets of genes that exhibit distinctive time patterns, visually encoded using colour. The dataset consist of the fold change of 800 differentially expressed genes in five time points. The original dataset can be found in the Gene Expression Omnibus (GEO) [EDL02] repository with accession number GSE49577 [KLHS14].*

the larger ones in which they often belong to. Hence, the second slider sets the maximum allowed similarity distance between a parent main cluster and a child sub-cluster. Distinctive *'weak-edges'* between neighbouring nodes are coloured *red*, to give a hint to the user. Experimenting with different 'distinctiveness' thresholds, can help in identifying possible outliers and nested clusters.

Each gene is represented as a line in the parallel coordinates view and as a rectangle at the dendrogram view. Similarity scores, which are always intermediate branch nodes, are represented as circles of diameters proportional to their value. The user can hover over those circles to get a preview of the branch-cut and then click to select it. Moreover, a gene of particular interest could be highlighted by double clicking on its mapped rectangle. Each selected subset could be then exported as a comma separated values (CSV) file.

4. Conclusion

As an alternative to approaches that make assumptions about the data in hand, we developed a visual analysis tool for the hierarchical clustering of any short time-series gene expression data. The tool makes the algorithm more transparent, by enabling a more steerable exploration of the dendrogram. Its interactive design can help in identifying nested clusters and outliers, through this synergistic approach.

References

[AMA08] ARCHAMBAULT D., MUNZNER T., AUBER D.: Grouseflocks: Steerable exploration of graph hierarchy space. *Visualization and Computer Graphics, IEEE Transactions on 14*, 4 (2008), 900–913. doi:10.1109/tvcg.2008.34. 1

[AS94] AHLBERG C., SHNEIDERMAN B.: Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1994), ACM, pp. 313–317. doi:10.1145/191666.191775. 2

[AVHK06] ABELLO J., VAN HAM F., KRISHNAN N.: Ask-graphview: A large scale graph visualization system. *Visualization and Computer Graphics, IEEE Transactions on 12*, 5 (2006), 669–676. doi:10.1109/tvcg.2006.120. 1

[CCKK12] CRAIG P., CANNON A., KUKLA R., KENNEDY J.: Matse: The microarray time-series explorer. In *Biological Data Visualization (BioVis), 2012 IEEE Symposium on* (2012), IEEE, pp. 41–48. doi:10.1109/biovis.2012.6378591. 1

[CMP09] CHEN J., MACEACHREN A. M., PEUQUET D. J.: Constructing overview+ detail dendrogram-matrix views. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 889–896. doi:10.1109/tvcg.2009.130. 2

[CRV08] CARDONA G., ROSSELLÓ F., VALIENTE G.: Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC bioinformatics 9*, 1 (Jan. 2008), 532. doi:10.1186/1471-2105-9-532. 2

[CZWT15] CHAN S.-C., ZHANG L., WU H.-C., TSUI K.-M.: A maximum a posteriori probability and time-varying approach for inferring gene regulatory networks from time course gene microarray data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on 12*, 1 (2015), 123–135. doi:10.1109/tcbb.2014.2343951. 2

[DCMK07] DOTAN-COHEN D., MELKMAN A. A., KASIF S.: Hierarchical tree snipping: Clustering guided by prior knowledge. *Bioinformatics 23*, 24 (2007), 3335–3342. doi:10.1093/bioinformatics/btm526. 1

[EBJ06] ERNST J., BAR-JOSEPH Z.: Stem: a tool for the analysis of short time series gene expression data. *BMC bioinformatics 7*, 1 (2006), 1. doi:10.1186/1471-2105-7-191. 2

[EC02] ESTIVILL-CASTRO V.: Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter 4*, 1 (2002), 65–75. doi:10.1145/568574.568575. 1

[EDL02] EDGAR R., DOMRACHEV M., LASH A. E.: Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research 30*, 1 (2002), 207–210. doi:10.1093/nar/30.1.207. 2

[HBW*14] HESS M., BREMM S., WEISSGRAEBER S., HAMACHER K., GOESELE M., WIEMEYER J., VON LANDESBERGER T.: Visual exploration of parameter influence on phylogenetic trees. *IEEE Computer Graphics and Applications 34* (2014), 48–56. doi:10.1109/MCG.2014.2. 2

[HS01] HOCHHEISER H., SHNEIDERMAN B.: Interactive exploration of time series data. In *Discovery Science* (2001), Springer, pp. 441–446. doi:10.1007/3-540-45650-3_38. 2

[ID91] INSELBERG A., DIMSDALE B.: Parallel coordinates. In *Human-Machine Interactive Systems*. Springer, 1991, pp. 199–233. doi:10.1007/978-1-4684-5883-1_9. 2

[KK99] KARYPIS G., KUMAR V.: Chameleon: hierarchical clustering using dynamic modeling. *Computer 32*, 8 (1999), 68–75. doi:10.1109/2.781637. 1

[KLHS14] KOUSSOUNADIS A., LANGDON S., HARRISON D., SMITH V. A.: Chemotherapy-induced dynamic gene expression changes in vivo are prognostic in ovarian cancer. *British journal of cancer 110*, 12 (2014), 2975–2984. doi:10.1038/bjc.2014.258. 2

[LZH08] LANGFELDER P., ZHANG B., HORVATH S.: Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England) 24*, 5 (Mar. 2008), 719–20. doi:10.1093/bioinformatics/btm563. 1

[MABK11] MAHANTA P., AHMED H., BHATTACHARYYA D., KALITA J. K.: Triclustering in gene expression data analysis: a selected survey. In *Emerging Trends and Applications in Computer Science (NCETACS), 2011 2nd National Conference on* (2011), IEEE, pp. 1–6. doi:10.1109/ncetacs.2011.5751409. 1

[MGT*03] MUNZNER T., GUIMBRETIÈRE F., TASIRAN S., ZHANG L., ZHOU Y.: Treejuxtaposer: scalable tree comparison using focus+ context with guaranteed visibility. In *ACM Transactions on Graphics (TOG)* (2003), vol. 22, ACM, pp. 453–462. doi:10.1145/1201775.882291. 2

[MR10] MCGUFFIN M. J., ROBERT J.-M.: Quantifying the space-efficiency of 2d graphical representations of trees. *Information Visualization 9*, 2 (2010), 115–140. doi:10.1057/ivs.2009.4. 2

[MWS*10] MEYER M., WONG B., STYCZYNSKI M., MUNZNER T., PFISTER H.: Pathline: A tool for comparative functional genomics. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 1043–1052. doi:10.1111/j.1467-8659.2009.01710.x. 2

[NWN*09] NAVLAKHA S., WHITE J., NAGARAJAN N., POP M., KINGSFORD C.: Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5541 LNBI*, 3 (Mar. 2009), 400–417. doi:10.1007/978-3-642-02008-7_29. 1

[RAHZ13] REDDY C. K., AL HASAN M., ZAKI M. J.: Clustering biological data. In *Data clustering: algorithms and applications*. CRC Press, 2013, pp. 381–413. 1

[SM11] SINHA A., MARKATOU M.: A platform for processing expression of short time series (pests). *BMC bioinformatics 12*, 1 (2011), 1. doi:10.1186/1471-2105-12-13. 1

[SS02] SEO J., SHNEIDERMAN B.: Interactively exploring hierarchical clustering results [gene identification]. *Computer 35*, 7 (2002), 80–86. doi:10.1109/mc.2002.1016905. 2

[WWLC08] WANG X., WU M., LI Z., CHAN C.: Short time-series microarray analysis: Methods and challenges. *BMC Systems Biology 2*, 1 (2008), 58. doi:10.1186/1752-0509-2-58. 1