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

Thanasis Vogogias[1] (t.vogogias@napier.ac.uk), Jessie Kennedy[1] (j.kennedy@napier.ac.uk), Daniel Archambault[2] (d.w.archambault@swansea.ac.uk)

[1]School of Computing, Edinburgh Napier University
[2]Department of Computer Science, Swansea University

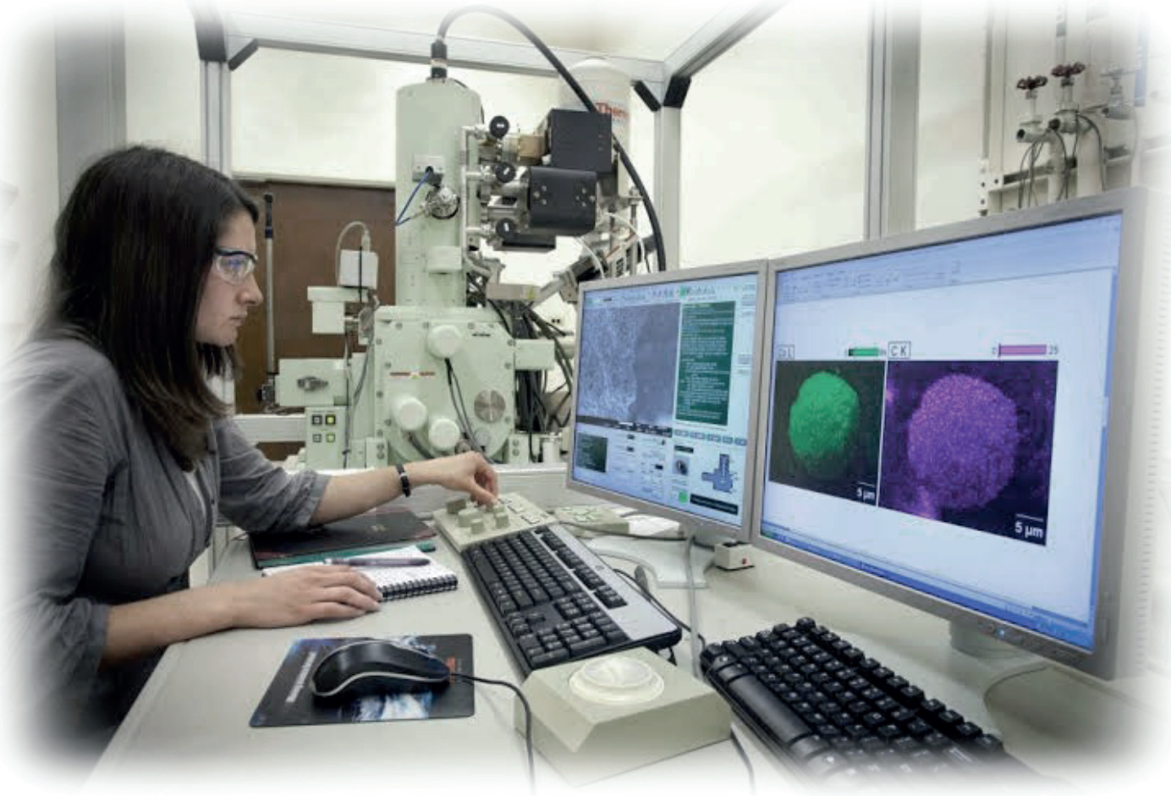
MOTIVATION

- There is an abundance of short time-series gene expression data, publicly available.

- Hierarchical clustering algorithms are used for their analysis and produce large dendrograms, which are hard to explore.

- Automated and semi-automated approaches make assumptions about the data.

- Alternatively, a more steerable approach could be followed.



EVALUATION

- A prototype has been developed in collaboration with biologists for analysing their own datasets.

- The prototype has been tested iteratively to refine features and capture new requirements.

- The intention was to support interactive multiple-height branch-cut.

Preprocessing microarray data

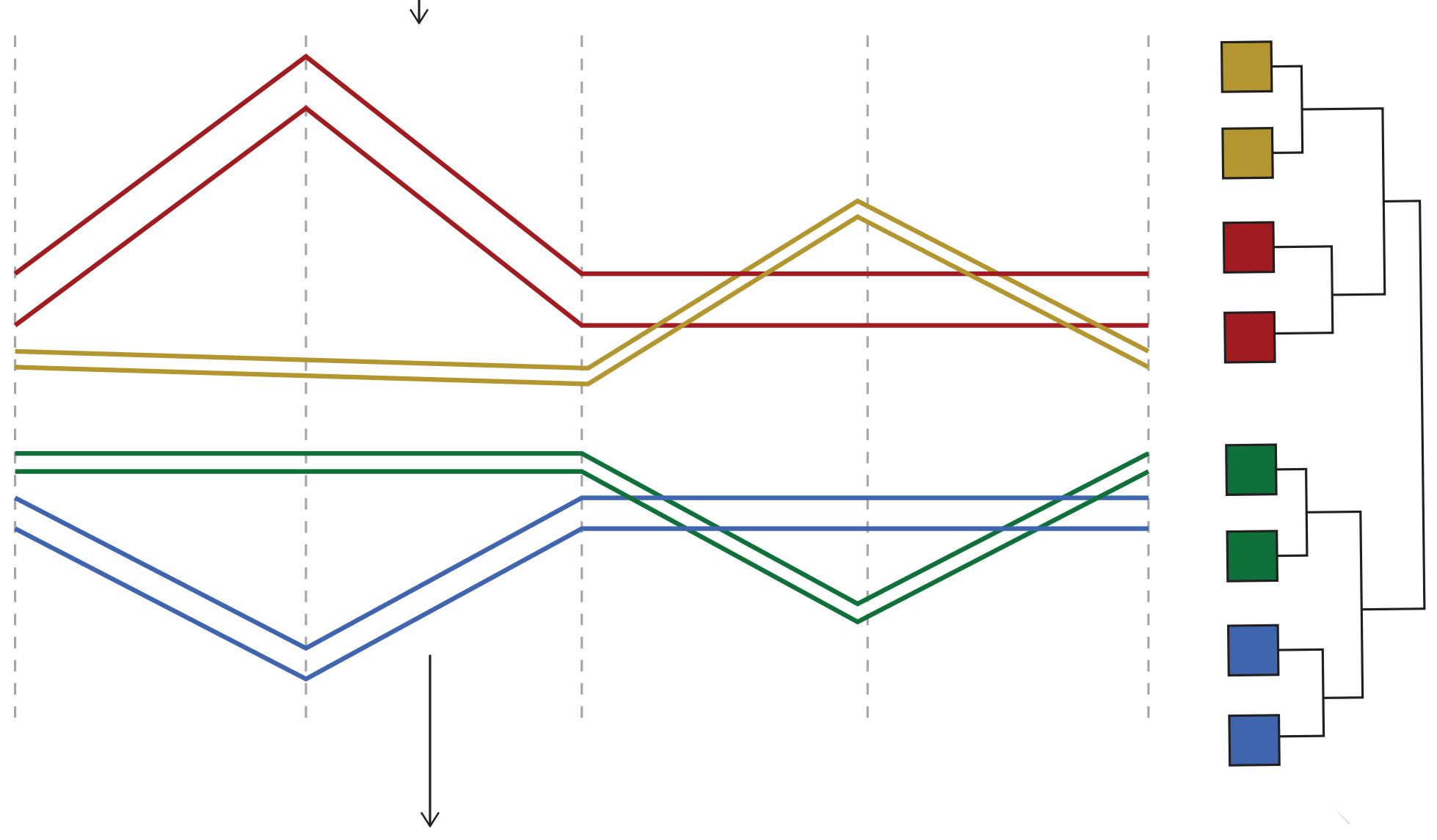
R Bioconductor { lumi
limma

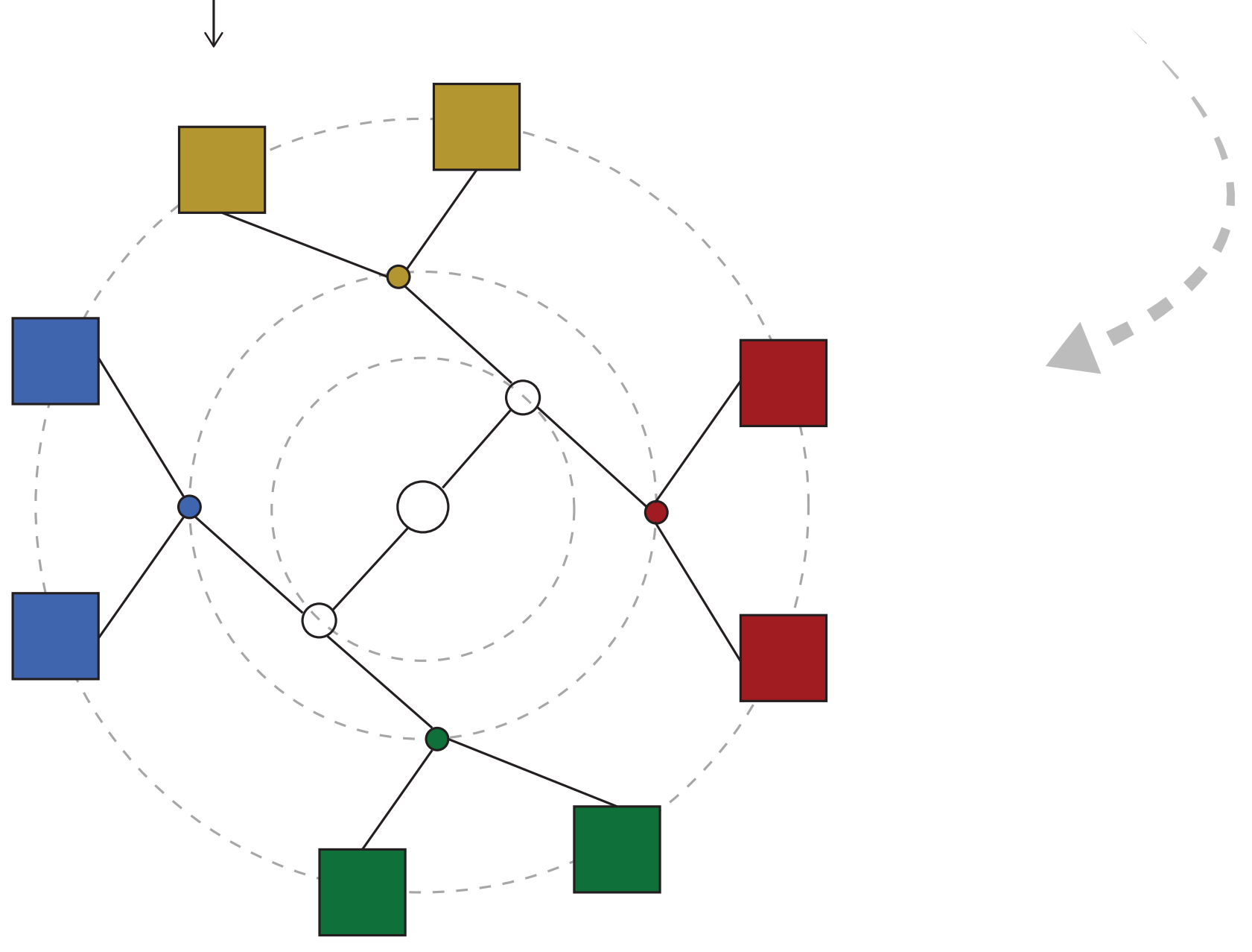
P-value < 0.05

Means of fold change of differentially expressed genes

NAME	DAY1	DAY2	DAY4	DAY7	DAY14
ILMN_2053546	-0.648248	0.027335	-0.03789	-0.840129	-0.173554
ILMN_1742981	0.596445	0.289452	-0.167664	0.170265	-0.053307
ILMN_3224758	0.51500	0.072121	-0.048392	0.063114	-0.103068
ILMN_1755115	-0.432101	0.044306	-0.086660	-0.540994	-0.237224
ILMN_1789702	0.009098	0.123383	0.215799	-1.275781	0.026091
ILMN_1742981	0.596445	0.289452	-0.167664	0.170265	-0.053307
ILMN_2053546	-0.648248	0.027335	-0.03789	-0.840129	-0.173554
ILMN_1755115	-0.432101	0.044306	-0.086660	-0.540994	-0.237224

PC (representation)





ANALYSIS & DESIGN

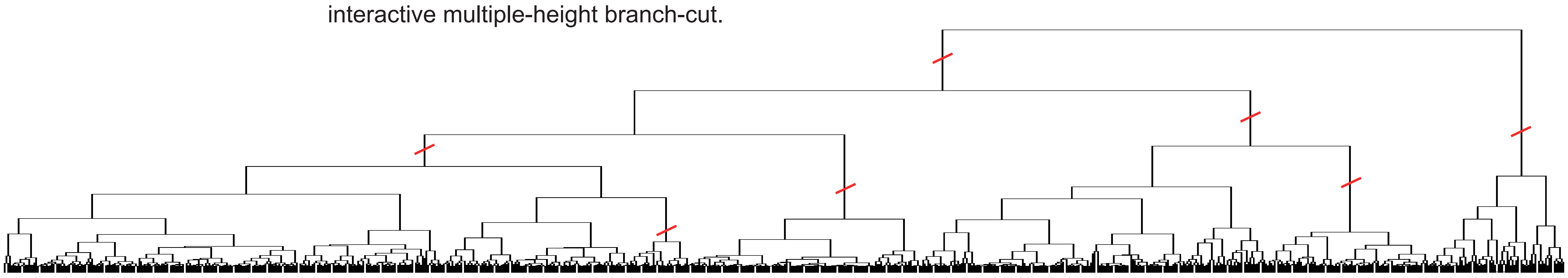
- Short time-series gene expression data are represented as parallel coordinates (PC).

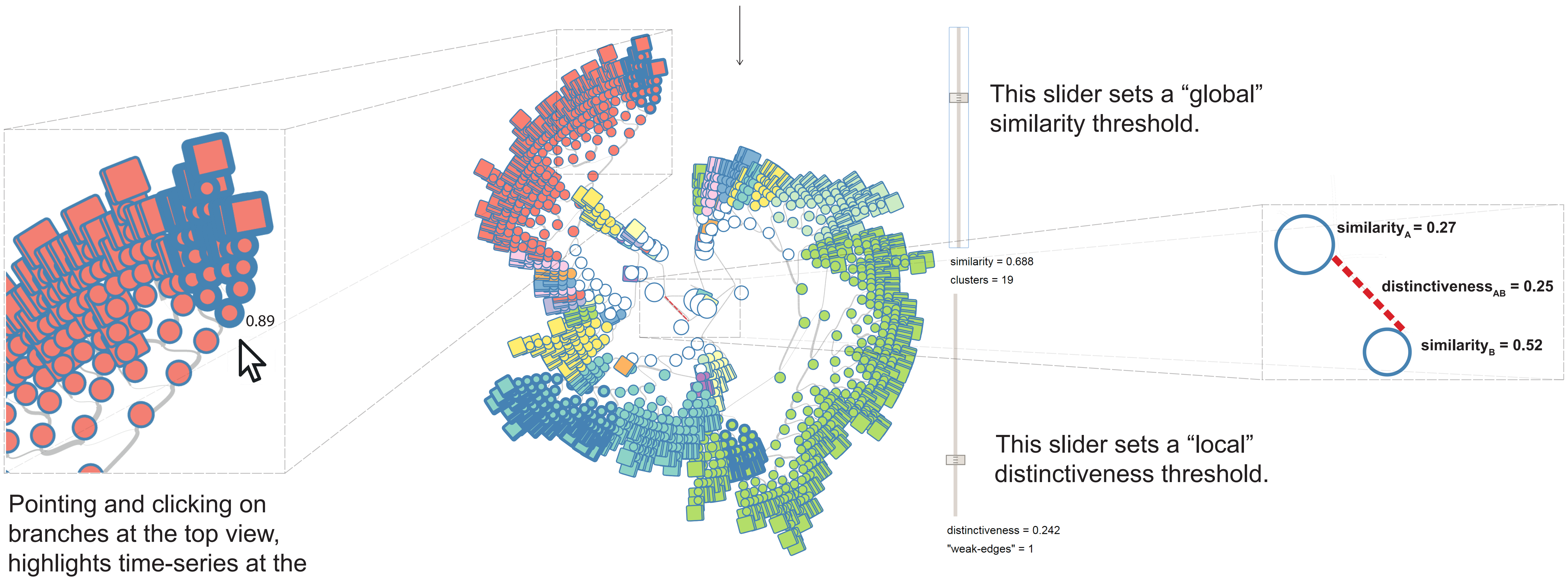
HCA ——— R { TSclust
hclust
dist

- A hierarchical clustering algorithm (HCA) is used to detect time patterns.

- The display of large dendrograms can be improved using a radial layout.

- The visual encoding follows perception principles and biological drawing conventions.

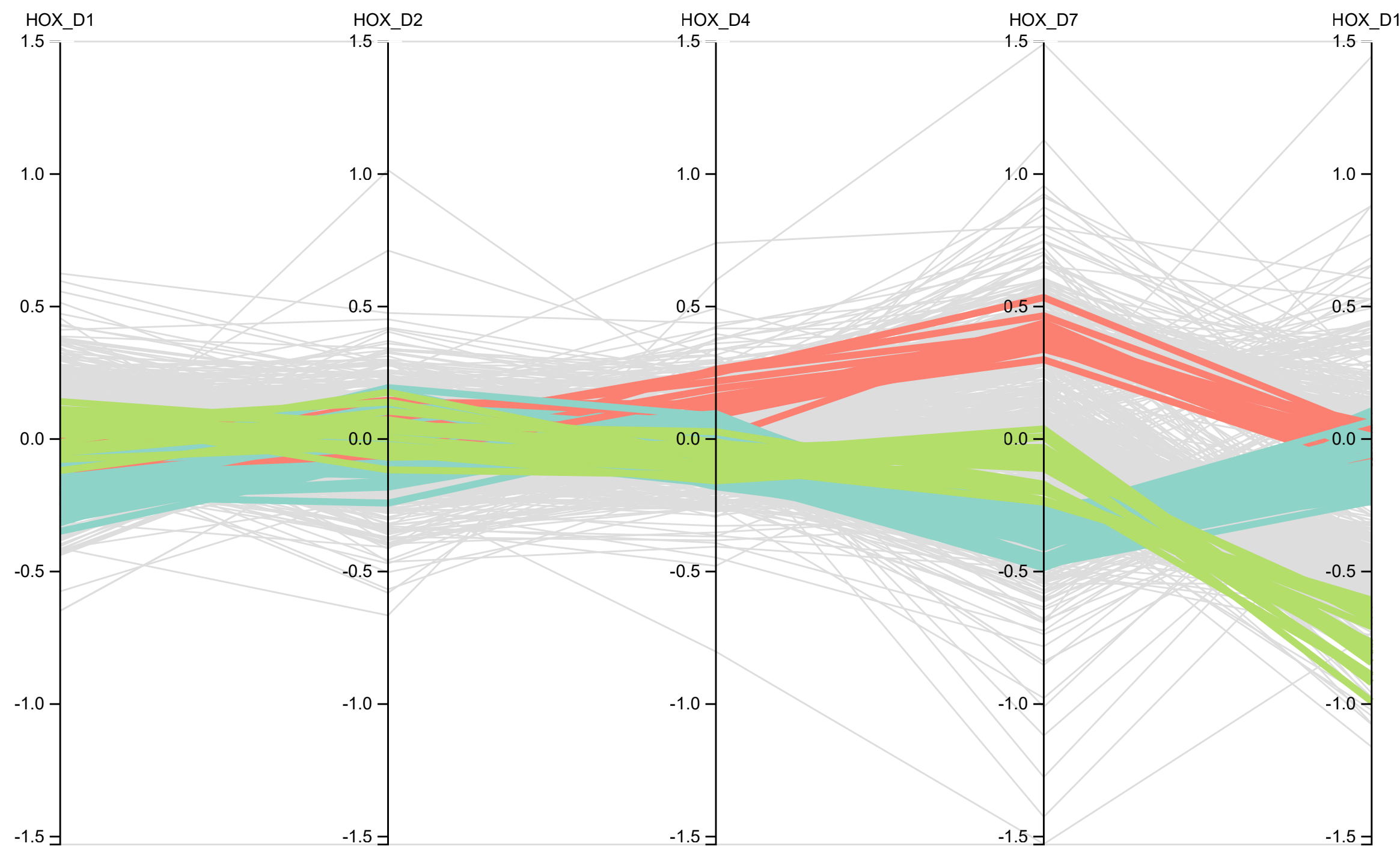




This slider sets a “global” similarity threshold.

This slider sets a “local” distinctiveness threshold.

Pointing and clicking on branches at the top view, highlights time-series at the linked bottom view.



CONCLUSION

- Hierarchical clustering algorithms are used to find patterns in short time-series gene expression data.

- However, the visual exploration of large dendrograms is problematic.

- Therefore, we developed a visual analytics approach for a more steerable exploration, that enables multiple-height branch-cut.