Coral: a suite of visualizations for comparing clusterings

Darya Filippova*, Aashish Gadani, Carl Kingsford {dfilippo | carlk @ cs.umd.edu, aagadani@umd.edu}
CBCB, Computer Science Department, University of Maryland
* - corresponding author

The presentation of this work was made possible by a Travel Fellowship awarded by ISCB with grant funds obtained from the DOE - Department of Energy Office of Science.

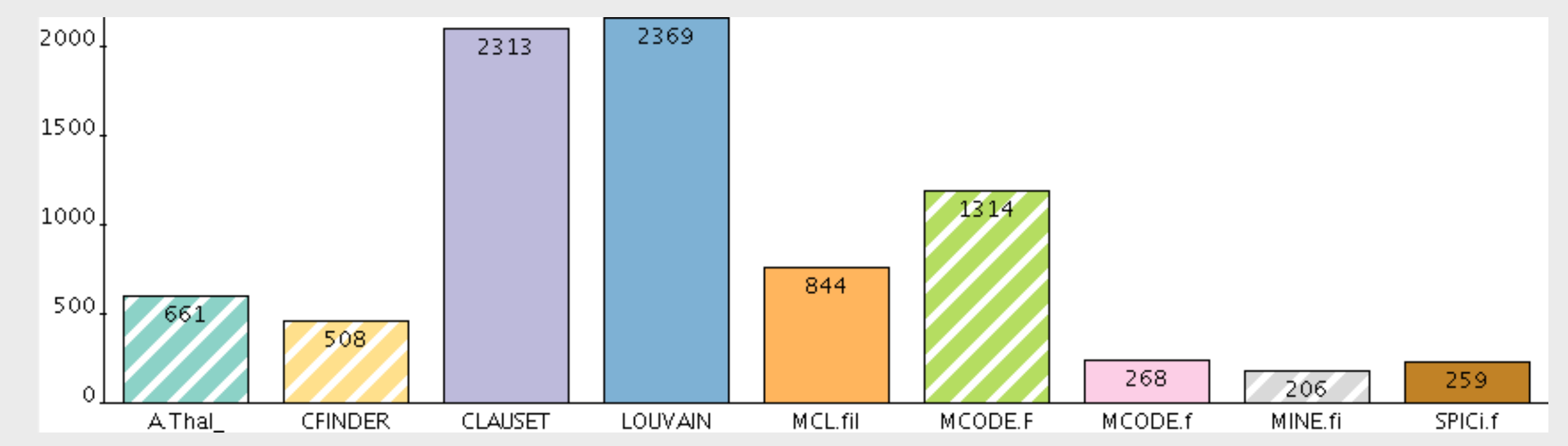
Multiple clusterings

Clustering has become a standard analysis for many types of biological data: interaction networks, gene expression, metagenomic abundance, genomic sequences. However, it is possible to obtain a large number of contradictory clusterings of the same data by:

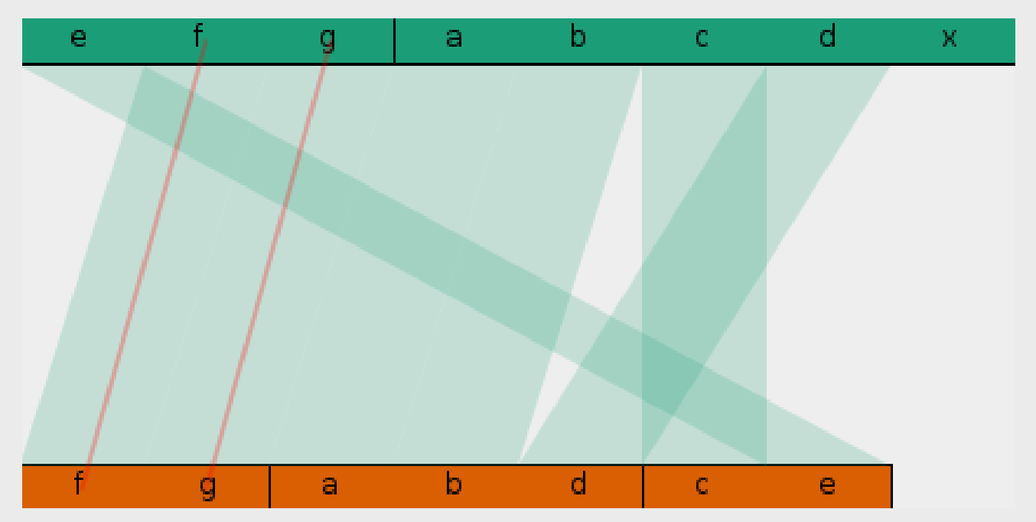
- **using different clustering algorithms [1-5]**
- **tuning algorithm parameters**
- **exploring optimal and near-optimal solutions [6, 7]**
- **clustering time-varying data**

Visual comparison

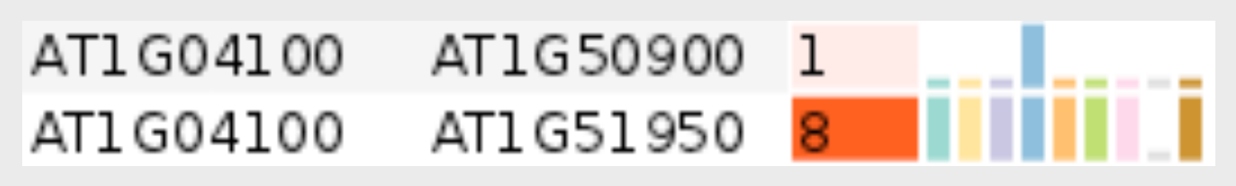
Visualization mantra: overview, zoom and filter, details on demand. Users can move from comparing all pairs of clusterings to examining individual item co-clustering patterns.
Coordinated displays: selecting an item in one view selects the corresponding items in other views.
Dataset statistics: gain a quick overview of the data by looking at a collection of bar charts for the number of modules, average module size, the number of data items covered by a clustering, or the percentage of elements in the overlapping modules.



Global similarity: explore clustering similarity using the ladder widget. White and pale blue cells represents pairs of disagreeing clusterings while bright blue cells highlight similarity.
Module comparison: narrow down the exploration to two clusterings and compare their modules. Users can filter the modules by Jaccard similarity and sort the table based on the size of the modules or their intersections and set differences.
Parallel partitions plot: track individual items and whole modules across multiple clusterings at a high levels. The bands of the plot are reordered to match the ordering of columns in the co-cluster matrix and to minimize the number of crossings.



Item pairs: a table visually encodes in which clusterings the two data items have been placed in a module together. Users can sort the columns to group pair with similar co-cluster patterns or search for patterns using the filter box.

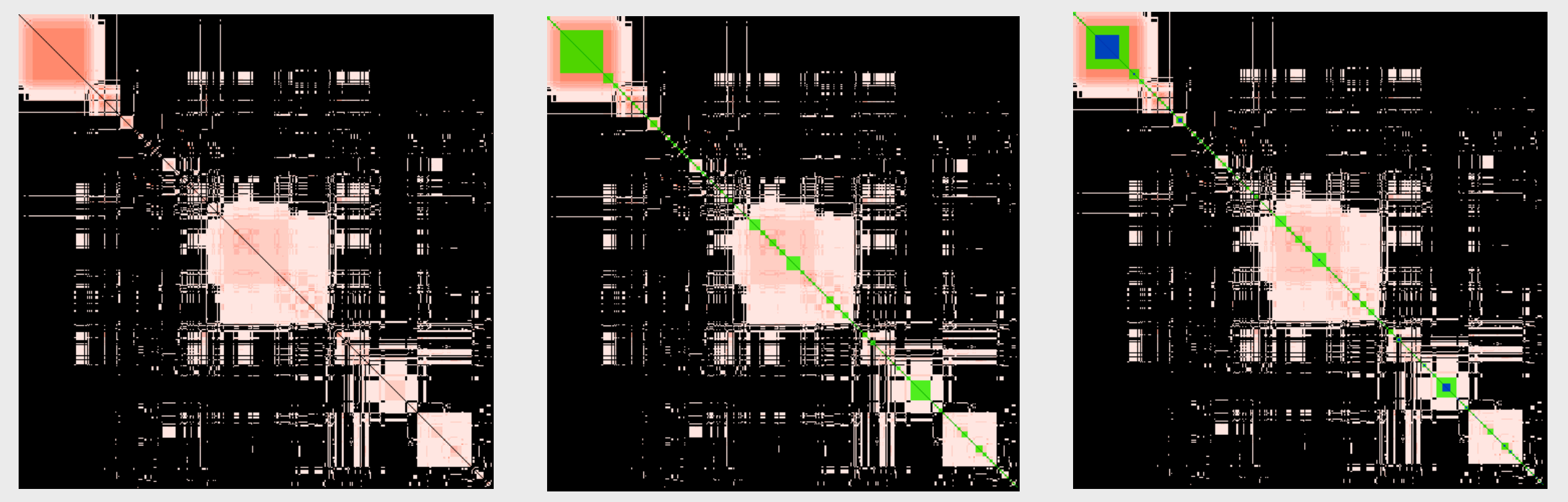


AT1G04100	AT1G50900	1
AT1G04100	AT1G51950	8

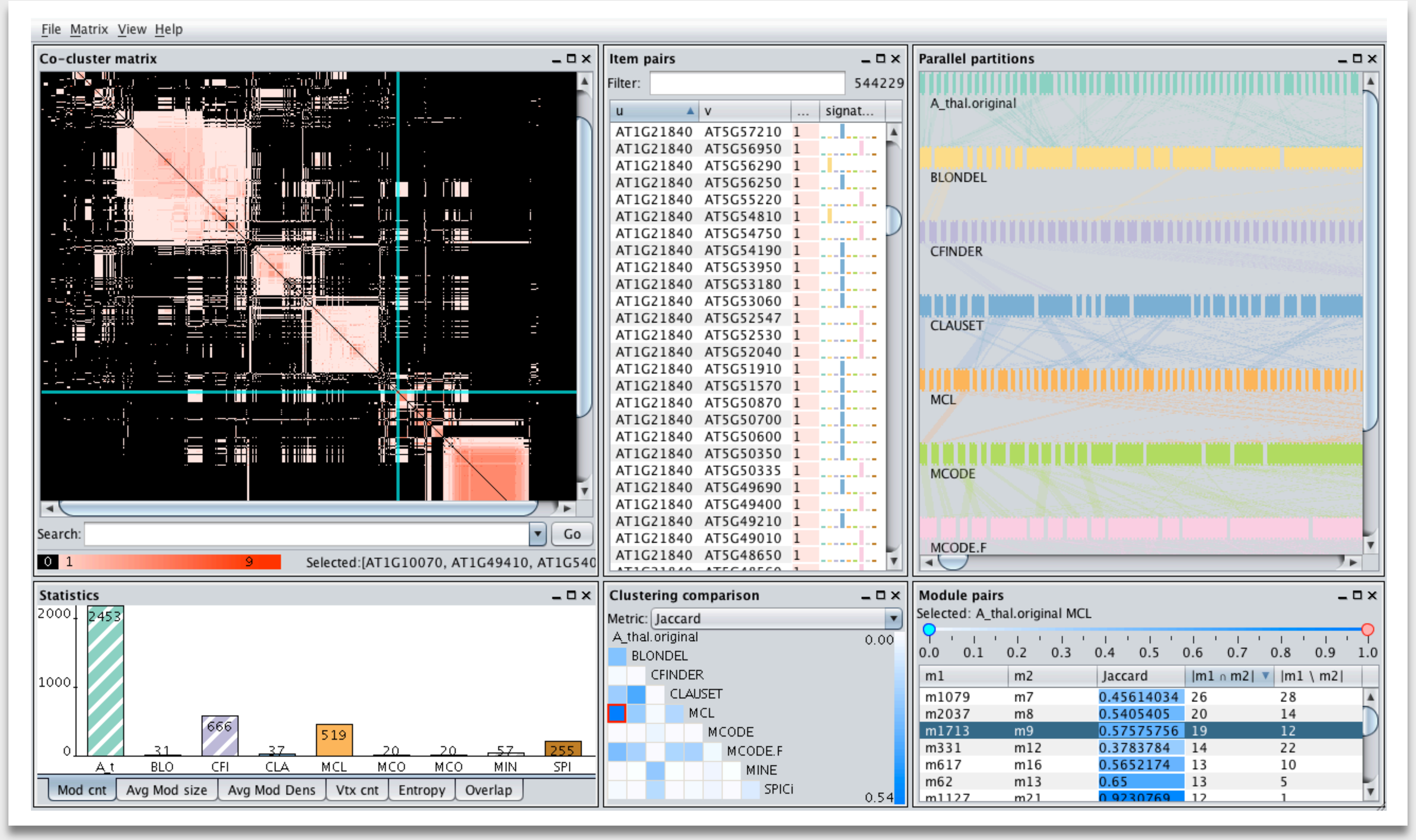
Co-cluster matrix: records which items were placed in the same module in a given clustering K. Matrix elements are defined as:

$$a_{ij}^{K} = \begin{cases} 0 & v_i \text{ and } v_j \text{ are in different modules in } K \\ 1 & v_i \text{ and } v_j \text{ are in the same module in } K. \end{cases}$$

We search for dense areas in the matrix using a dynamic programming approach. We call such dense areas *cores* (highlighted in green). If the original data was a network, we show the core's *cohesion* — a measure of separation of a core from the rest of the network (highlighted in blue).



Coral



Selected:[AT1G10070, AT1G49410, AT1G540

Filter: 544229

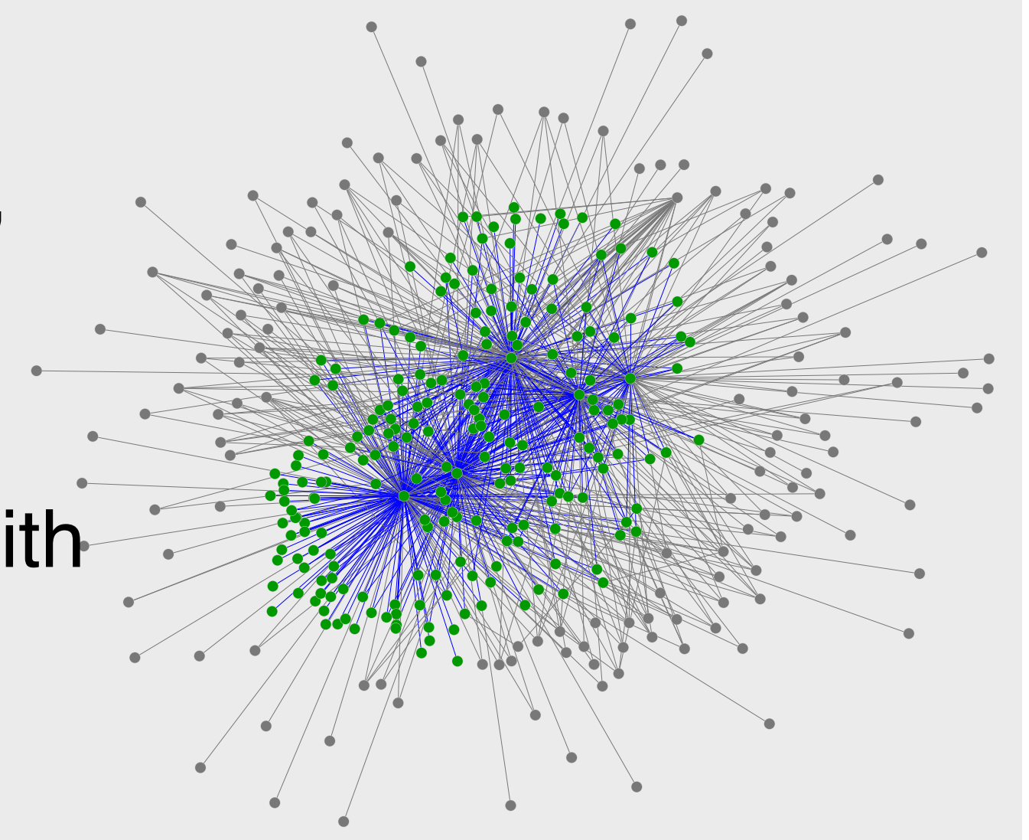
u	v	...	signat...
AT1G21840	AT5G57210	1	
AT1G21840	AT5G56950	1	
AT1G21840	AT5G56290	1	
AT1G21840	AT5G56250	1	
AT1G21840	AT5G55220	1	
AT1G21840	AT5G54810	1	
AT1G21840	AT5G54750	1	
AT1G21840	AT5G54190	1	
AT1G21840	AT5G53950	1	
AT1G21840	AT5G53180	1	
AT1G21840	AT5G53060	1	
AT1G21840	AT5G52547	1	
AT1G21840	AT5G52530	1	
AT1G21840	AT5G52040	1	
AT1G21840	AT5G51910	1	
AT1G21840	AT5G51570	1	
AT1G21840	AT5G50870	1	
AT1G21840	AT5G50700	1	
AT1G21840	AT5G50600	1	
AT1G21840	AT5G50350	1	
AT1G21840	AT5G50335	1	
AT1G21840	AT5G49690	1	
AT1G21840	AT5G49400	1	
AT1G21840	AT5G49210	1	
AT1G21840	AT5G49010	1	
AT1G21840	AT5G48650	1	

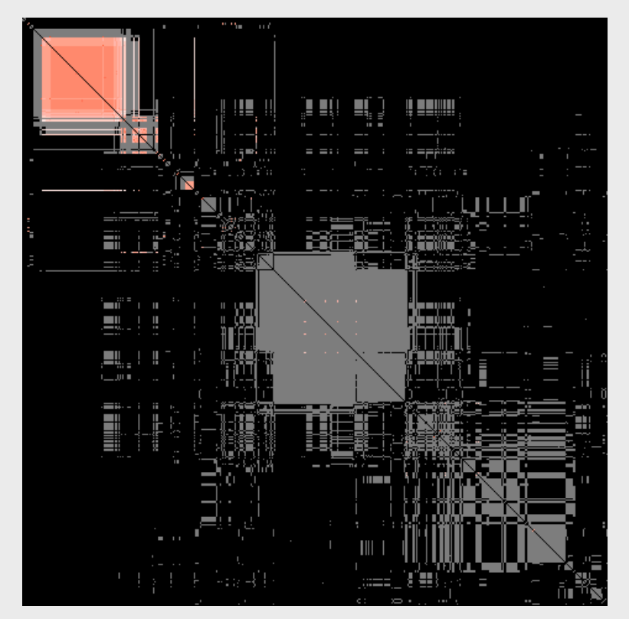
Selected: A_thal.original MCL

m1	m2	Jaccard	|m1 ∩ m2|	|m1 \ m2|
m1079	m7	0.45614034	26	28
m2037	m8	0.5405405	20	14
m1713	m9	0.57575756	19	12
m331	m12	0.3783784	14	22
m617	m16	0.5652174	13	10
m62	m13	0.65	13	5
m1127	m21	0.9230769	12	1

Example: *A. thaliana PPI network*

We applied 9 different clustering algorithms to a recent *A. thaliana* protein-protein interaction network [7]. Modules of size < 6 and with *partition density* < 0 were removed from the clusterings (as described in [7]).
Most cells in the ladder widget (left) are pale blue or white indicating that clusterings **mostly disagreed**.
The jaccard similarity for the MCL [4] and link clustering [3] pair can be explained via the module-to-module table: both clusterings identified a large module of 288 proteins and had exact matches among several smaller modules.
65.25% of all items in the co-cluster matrix have the value of 1, and only 6.34% of the matrix' cells have the value of 5 or greater.
The DP identified 249 cores, most of them with a low cohesion, i.e. proteins in the cores shared many edges with proteins outside of the core (see left column, bottom).





Setting the link clustering [3] as a *base* reveals that link clustering recovered only a fraction of modules found by other algorithms, but did include the largest module.

References

1. Adamcsek et al. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 2006, 22(8):1021–3.
2. Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks. BMC *Bioinformatics* 2003, 4:2
3.Ahn Y, et al. Link communities reveal multiscale complexity in networks. *Nature Letters* 2010, 466(5):761–765.
4. van Dongen S. Graph clustering by flow simulation. *PhD thesis*, University of Utrecht 2000.
5. Jiang P, Singh M. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics* 2010, 26(8):1105–11.
6. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008, 24(5):719–720
7. Lewis A.C., et al. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology* 2010, 4:100.
8. Arabidopsis Interactome Mapping Consortium. Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science* 2011, 333(6042):601–607.

Download at: www.cbcb.umd.edu/kingsford-group/coral