



Data Mining w/o Programming

A hands-on workshop at the
Functional Genomics Workshop, Ljubljana, Slovenia

These notes include Orange workflows that we will construct, and visualizations we will create during the workshop.

Workshop instructors:

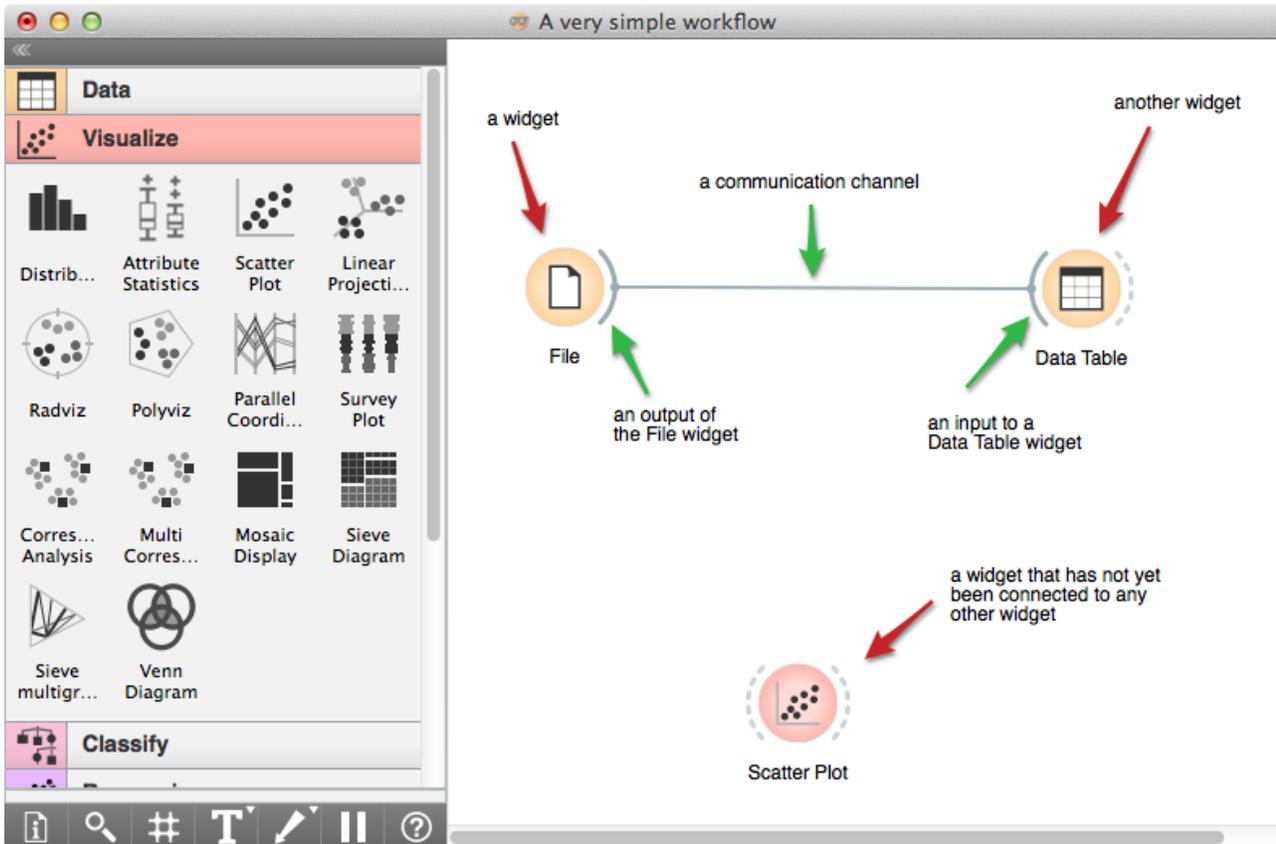
Blaz Zupan, Janez Demsar and Tomaž Curk, with help from members of Bioinformatics Lab, Ljubljana.

Welcome to the hands-on Data Mining workshop! This three-hour workshop is designed for students and researchers in molecular biology. You will see how common data mining tasks can be accomplished without programming. We will use Orange to construct visual data mining flows. Many similar data mining environments exist, but the organizers prefer Orange for a simple reason—they are its authors.

If you haven't already installed Orange, please follow the installation guide at <http://biolab.github.io/functional-genomics-workshop-orange>

Lesson 1: Workflows in Orange

Orange workflows consist of components that read, process and visualize data. We call them “widgets”. Widgets are placed on a drawing board—the “canvas”. Widgets communicate by sending information along a communication channel. Output from one widget is used as input to another.

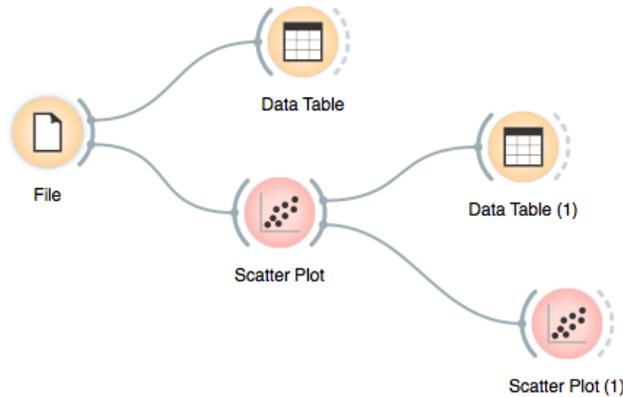


A simple workflow with two connected widgets and one widget without connections. The outputs of a widget appear on the right, while the inputs appear on the left.

We construct workflows by dragging widgets onto the canvas and connecting them by drawing a line from the transmitting widget to the receiving widget.

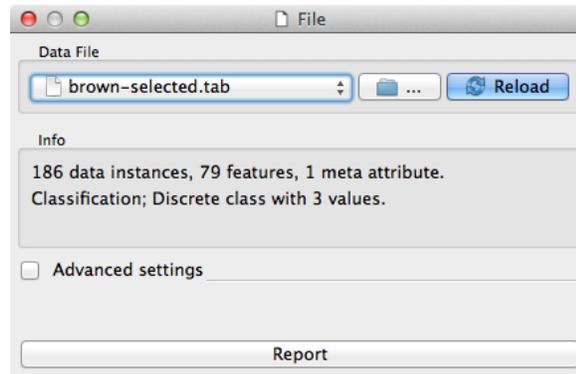
Construct a data flow that consists of a File widget, two Scatter Plot widgets and two Data Table widgets.

Workflow with a File widget that reads data from disk and sends it to the Scatter Plot and Data Table widget. The Data Table renders data in a spreadsheet, while the Scatter Plot visualizes it. Selected data points from the Scatterplot are sent to two other widgets: Data Table (1) and Scatter Plot (1).



The File widget reads data from disk. Open the File Widget by double clicking its icon. Orange comes with several preloaded data sets. From these (“Browse documentation data sets...”), choose brown-selected.tab, a yeast gene expression data set.

Orange workflows often start with a File widget. The brown-selected data set comprises 186 rows (genes) and 81 columns. Out of the 81 columns, 79 contain gene expressions of baker’s yeast under various conditions, one column (marked as a “meta attribute”) provides gene names, and one column contains the “class” value or gene function.

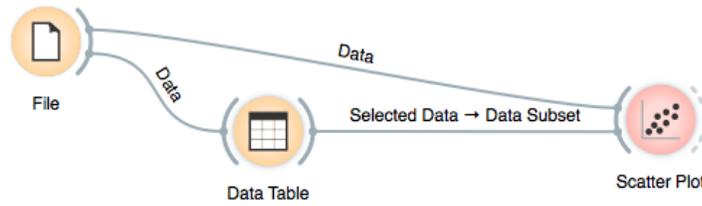


After you load the data, open the other widgets. In the Scatter Plot widget select a few data points and watch as they appear in Data Table (1). Use a combination of two Scatter Plot widgets, where the second scatterplot shows a detail of a smaller region selected in the first scatterplot.

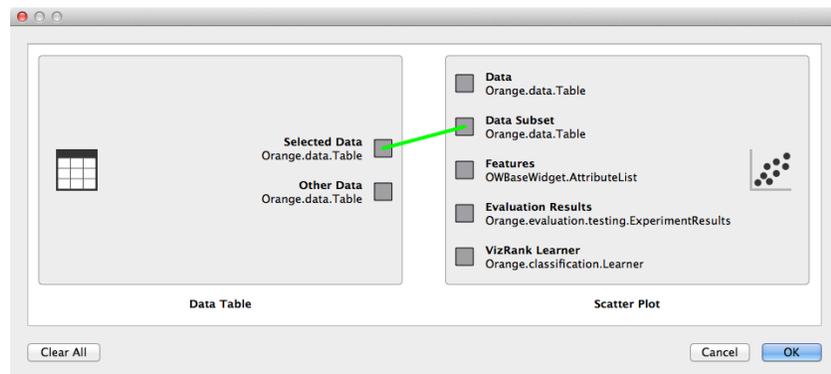
The Scatterplot for a pair of random features does not provide much information on gene function. Does this change with a different choice of the features? Try intelligent visualization scoring by VizRank, which is implemented within the Scatter Plot widget.

We can connect the output of the Data Table widget to the Scatter Plot widget to highlight chosen data instances (rows) in the scatterplot.

In this workflow we have switched on the option “Show channel names between widgets”.



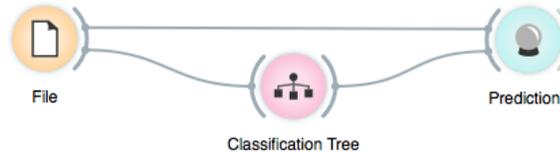
How does Orange distinguish between the primary data source and the selection? It uses the first connected signal as the entire data set and the next one as its subset. To make changes, double click on the line connecting the two widgets.



Lesson 2: Classification

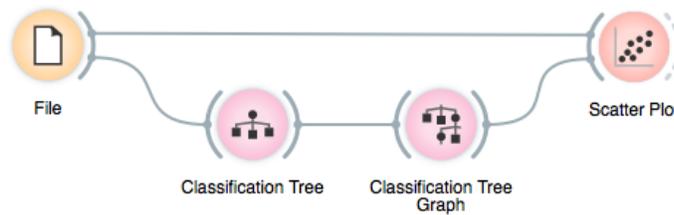
Genes in the yeast data set are labeled with three functions (“Proteas”, “Resp”, and “Ribo”). Can we construct a model that predicts the gene function based on the gene’s expression profile? We’ll first create a classification tree and observe its predictions.

Something in this workflow is conceptually wrong. Can you guess what?



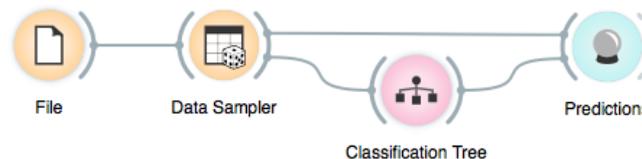
Classification trees split the data into smaller and smaller data sets until one of the classes prevails. We can use the Classification Tree Graph widget to visualize a classification tree model. Consider a combination with a scatterplot to visualize how the classification tree splits the data.

The Classification Tree widget outputs a classification tree model that is sent to the Classification Tree Graph widget, which renders the tree. Selecting a tree node in this widget will output the corresponding data.



In the next workflow we split the data set into two subsets: a training set and a test set. We construct the model from the training set, and observe the predicted class probabilities on the test set. Are the predictions reasonable? How can we assess their quality?

Widgets may transmit several types of signals. Data Sampler outputs both sampled data and left-out data. Orange will ask you which type of signal to pass to the receiving widget if it cannot resolve this automatically by matching the signal types.

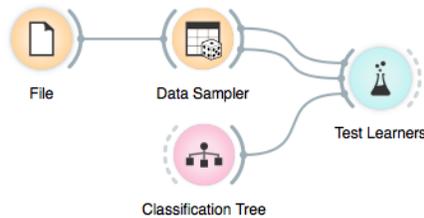


To observe which data instances were selected, feed the output of the Data Sampler widget to the Data Table or Info widgets.

Lesson 3: Classification Accuracy

To measure the quality of the constructed model we split the data into a training set and a test set. We evaluate the accuracy on data instances that have not been used for training. Accuracy can be measured by the proportion of data instances for which the class prediction was correct.

Try changing the size of the training set and observe the impact on accuracy. What do you expect? Try this with other data sets that come with Orange.



A predictor with 90% accuracy might sound good, but if 95 % of instances belong to the same class, it is actually worse than always predicting the majority class. For datasets with a skewed class distribution, other evaluation scores (such as Area Under ROC) are more appropriate.

Test Learners

Sampling

- Cross-validation
 - Number of folds: 5
- Leave-one-out
- Random sampling
 - Repeat train/test: 10
 - Relative training set size: 70%
- Test on train data
- Test on test data
- Apply on any change

Performance scores

- Classification accuracy
- Sensitivity
- Specificity
- Area under ROC curve
- Information score

Target class

Proteas

Evaluation Results

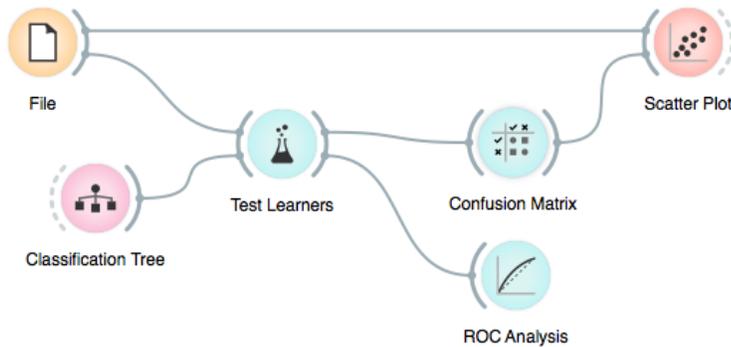
Method	CA	Sens	Spec	AUC	Brier
1 Classification Tree	0.9565	1.0000	0.9444	1.0000	0.0408

Report

Lesson 4: Cross-Validation

Estimating the accuracy may depend on a particular split of the data set. To increase robustness we can repeat the measurement several times, each time choosing a different subset of the data for training. One such method is cross-validation. It is available in Orange through the Test Learners widget. We will analyze its output by examining the confusion matrix and the ROC curve.

The Confusion Matrix widget outputs data instances related to the selected cells. In this schema we visualize them in the Scatter Plot widget as a data subset. What can you say about the misclassified instances? Does the scatterplot provide insights? Are there outliers?



In cross-validation each data instance is used for testing exactly once.

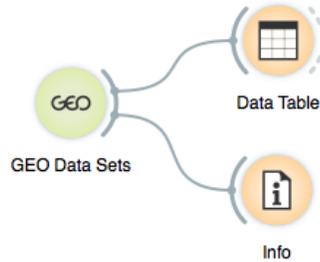
We can use the Confusion Matrix widget to find how many test instances were classified correctly and, if not, which class they were mistaken for.

	Proteas	Resp	Ribo	
Proteas	32	2	1	35
Resp	1	28	1	30
Ribo	0	0	121	121
	33	30	123	186

Lesson 5: GEO Data Sets

The bioinformatics add-on provides access to a data set library by Gene Expression Omnibus (GEO). Orange queries GEO for each selected data set and downloads it. Construct the depicted workflow and inspect a few data sets.

In the GEO Data Sets widget try changing the setting of what data will be represented in rows. Check the output in the Data Table and Info widgets. Which setting would be appropriate for creating a data set for classification?



The data sets that have been downloaded are marked with a bullet in the first column of the table.

Info
3269 datasets
3 datasets cached

Output
Rows
 Genes or spots
 Samples
 Merge spots of same gene
Data set name
Smoking effect on B lympho...

Commit

 Commit on any change

Filter

ID	Title	Organism	Samples	Features	Genes	Subsets	PubMedID
• GDS360	Breast cancer and docetaxel...	Homo sapiens	24	12625	9459	2	
• GDS3713	Smoking effect on B lympho...	Homo sapiens	79	22283	14047	2	20217071
• GDS1210	Gastric cancer	Homo sapiens	30	7129	6172	2	11782383
GDS2526	c-MYC depletion effect on c...	Homo sapiens	18	54675	31396	8	17159920
GDS2524	Effect of gonadal steroids o...	Mus musculus	48	22690	13916	10	16714546
GDS2525	Foxp3 ablation effect on m...	Mus musculus	4	45101	26722	2	17220892
GDS2522	Pyocyanin treatment: dose r...	Saccharomyc...	6	9335	8714	5	17185230

Description
Analysis of peripheral circulating B cells from smoking and non-smoking healthy US white females. B cells are directly associated with the onset and development of many smoking-induced diseases. Results provide insight into the molecular basis of B cell involvement in smoking-related pathogenesis.

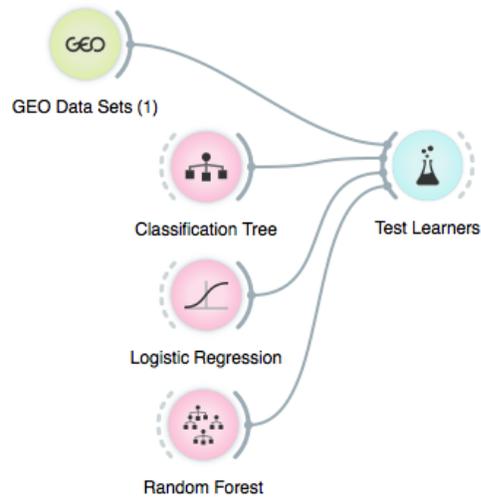
Sample Annotations

Type (Sample annotations)	Sample count
<input checked="" type="checkbox"/> stress	
<input checked="" type="checkbox"/> control	40
<input checked="" type="checkbox"/> cigarette smoke	39

Lesson 6: GEO Data Sets and Classification

From the GEO widget, select the data on breast cancer (GDS360) with 14 treatment resistant and 10 treatment sensitive tumors. Can we predict the treatment sensitivity from gene expression profiles?

The Random Forest classifier often achieves good accuracy on gene expression data. Try changing the number of classification trees in the forest. How does the accuracy change? Does random forest beat a single classification tree? How does logistic regression compare with the other two methods?

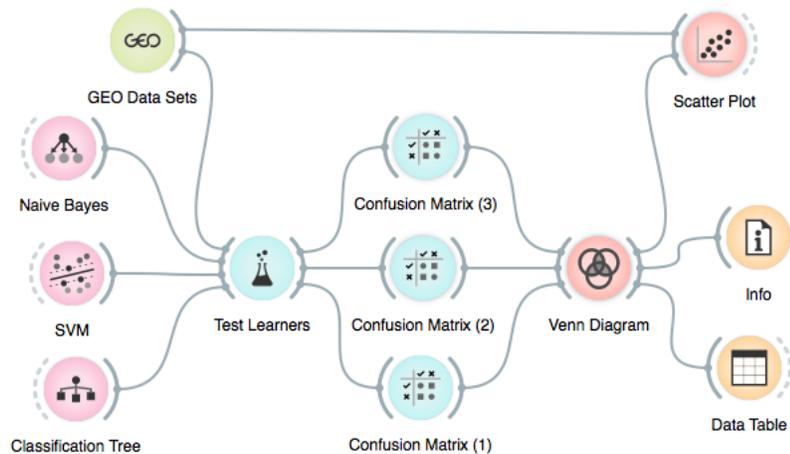


We will test the accuracy of three learners: classification tree, logistic regression, and random forest. We recommend starting with smaller data sets as some of the learning algorithms require a lot of time.

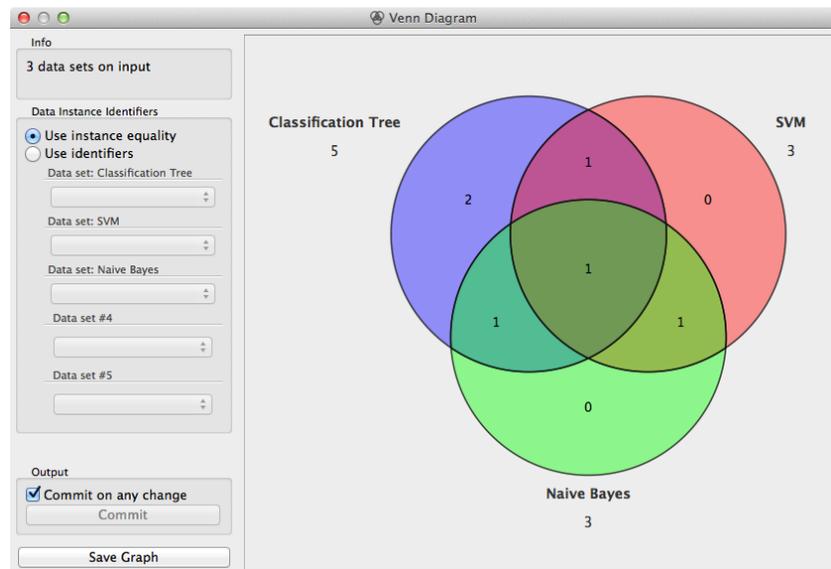
Lesson 7: Venn Diagram

The following workflow looks intimidating, but it's not as complicated as it looks. The question we are trying to answer is: do different classifiers misclassify the same tissue samples? That is, are some specific test instances hard to classify? Are they outliers, or even originally misclassified tissue samples? We can answer all but the last question by cross-validating the classifiers, selecting misclassified instances in the Confusion Matrix, and relating the three sets of misclassifications in the Venn diagram.

Most widgets in Orange are interactive. For example, you can click on different sections of the Venn diagram to output a related data item and inspect it with other widgets.



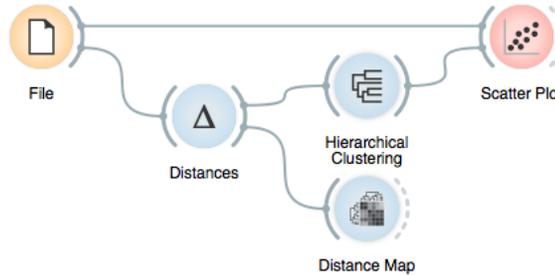
We can now choose various sections of the Venn Diagram and inspect which of the data instances were the hardest to classify.



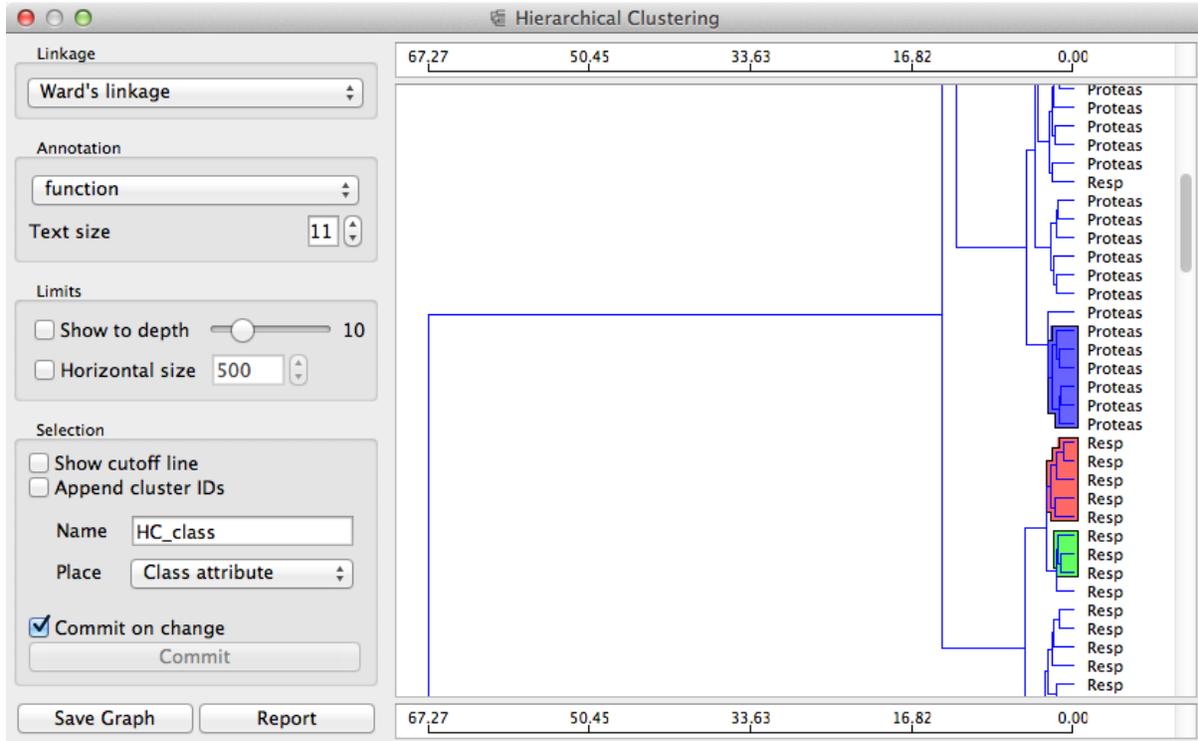
Lesson 8: Hierarchical Clustering

For hierarchical clustering, we need to measure the distances between genes (rows), which are fed into a Hierarchical Clustering widget that displays the dendrogram. The dendrogram is interactive: clicking on any branch sends its data instances to the output.

We used Euclidean distance (in the Distances widget) and Ward's linkage (in the Hierarchical Clustering widget). Euclidean distance may not be the best choice in this case. Do you agree? Experiment with other distance measures. Do you notice any changes in the dendrogram?



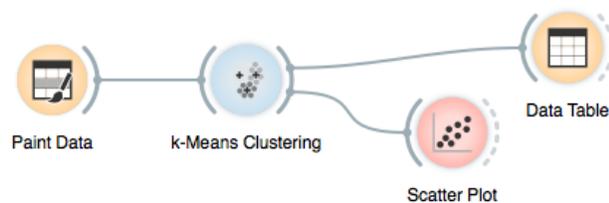
We display data instances selected in the dendrogram in a scatterplot. Make sure this widget is showing an informative visualization.



Lesson 9: k-Means Clustering

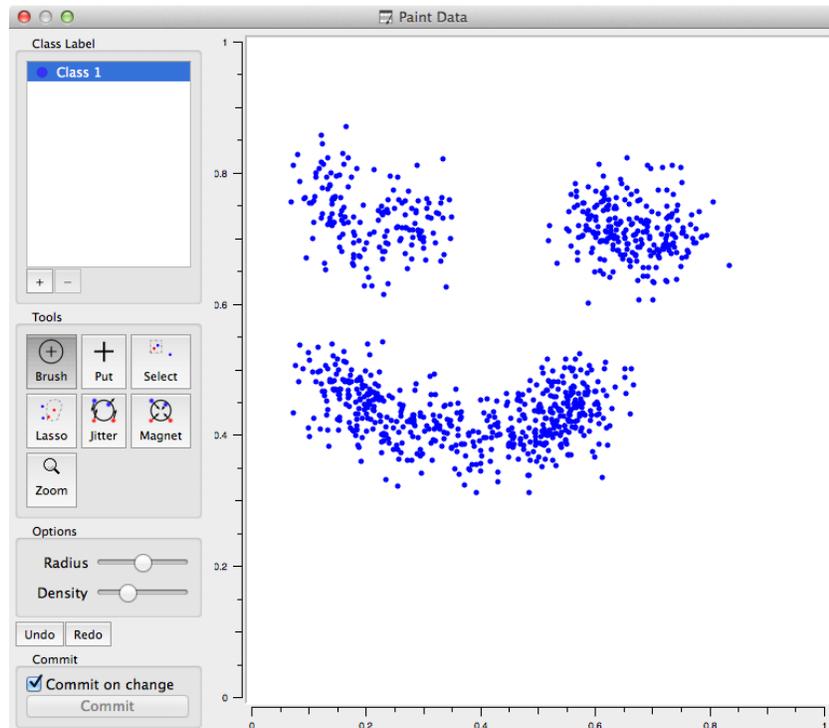
Hierarchical clustering is not suitable for larger data sets due to the prohibitive size of the distance matrix. An alternative approach, which doesn't use the distance matrix, is k-means clustering. Here we have to provide the number of clusters in advance.

Alternatively, we can use cluster scoring techniques to discover the optimal value for the number of clusters from a predefined range. You are free to try k-means clustering on any data set, however we will discuss its properties on hand-painted data.



A game we like to play is to see if silhouette scoring in k-means can discover the “correct” number of clusters.

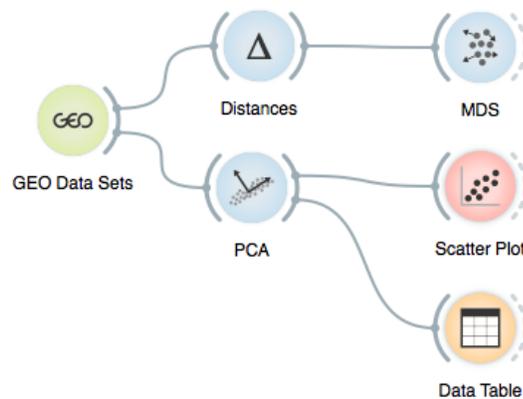
How many clusters do you see in the data set on the right? What is the number of clusters proposed by the silhouette method and k-means clustering? Help k-means find the expected number of clusters by modifying the data set.



Lesson 10: Data Projection

We have already seen one type of data projection—the scatterplot—but we were limited to projecting the data onto a hyperplane defined by two features. A technique that finds projections that retain the most variance is Principal Component Analysis (PCA). Another approach is Multidimensional Scaling (MDS), where we embed the data into a low dimensional space while trying to preserve distances between objects. The two approaches often yield similar visualizations.

Try replacing the GEO Data Sets widget with the File widget and select the brown-selected.tab dataset. Are the visualizations by PCA and MDS similar?

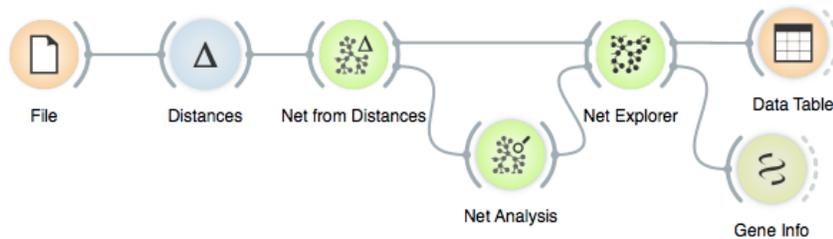


PCA can also be used for preprocessing by transforming the data to a lower dimensional space. This could sometimes increase accuracy, but also make the results harder to interpret.

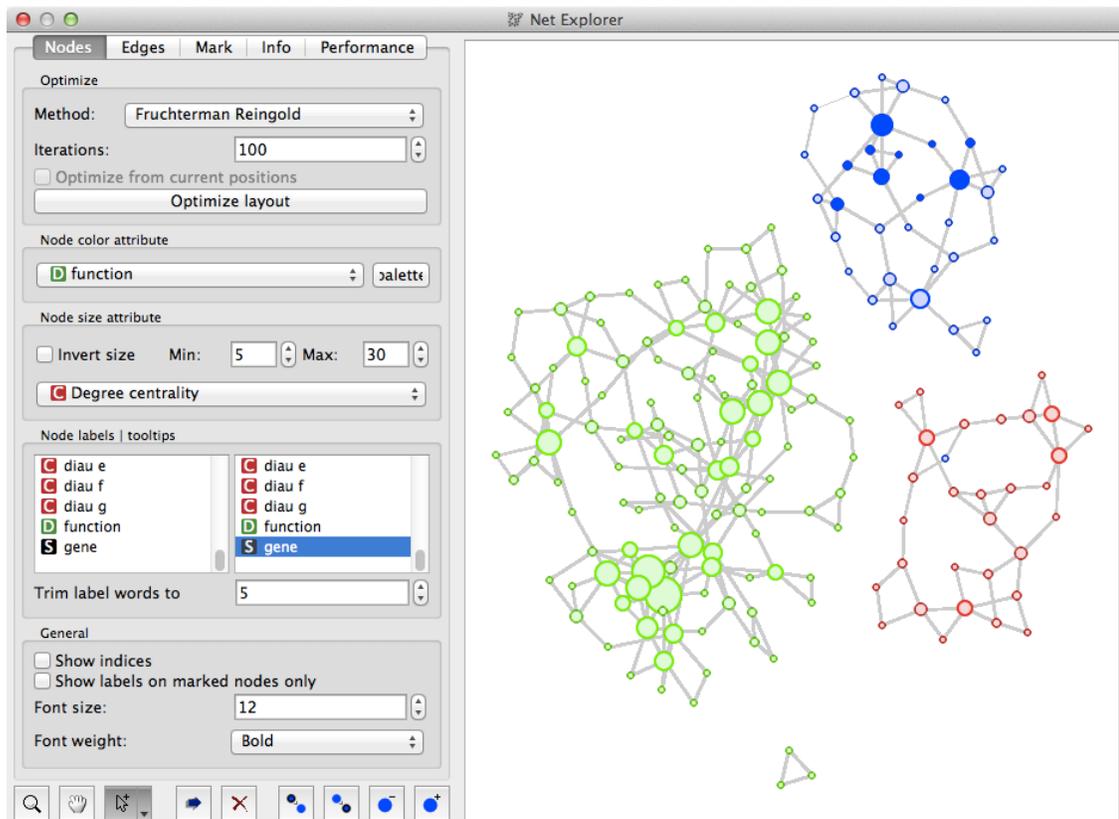
Lesson II: Correlation Networks

Similarity between data instances (e.g. genes, tissue samples, chemicals) can also be visualized with a network. We need to choose a similarity threshold or limit the number of edges per node. You need to have the Orange network add-on installed to construct and explore similarity-based networks.

Widgets in the network add-on provide many different options for visualization and analysis. How do the resulting networks change with different distance metrics? Are hubs invariant to the choice of the distance metric? Which are the hub genes?



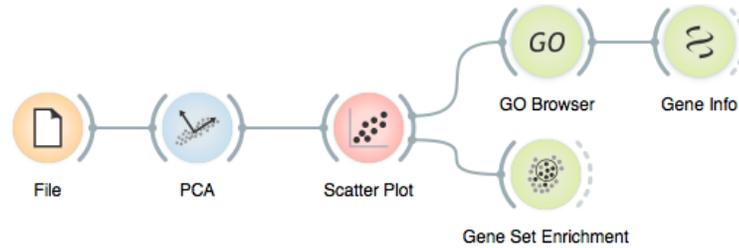
We added the Net Analysis widget to compute graph and node level statistics and pass them to the Net Explorer widget to be rendered in the network.



Lesson 12: Gene Set Enrichment

Data sets can store gene profiles in rows and also include gene names. We can use Orange workflows to select data instances, and see if the corresponding genes are present in some pathways or Gene Ontology terms. For this task the Orange bioinformatics add-on includes GO Browser and Gene Set Enrichment widgets.

Lists of gene sets (pathways, GO terms) in enrichment analysis widgets are clickable. Try rendering the output of these widgets in the Gene Info widget, and use it to find your favorite gene in the NCBI Gene database.



GO Browser presents two views of enriched pathways: one displaying the ontology tree and the other showing a list of enriched GO terms.

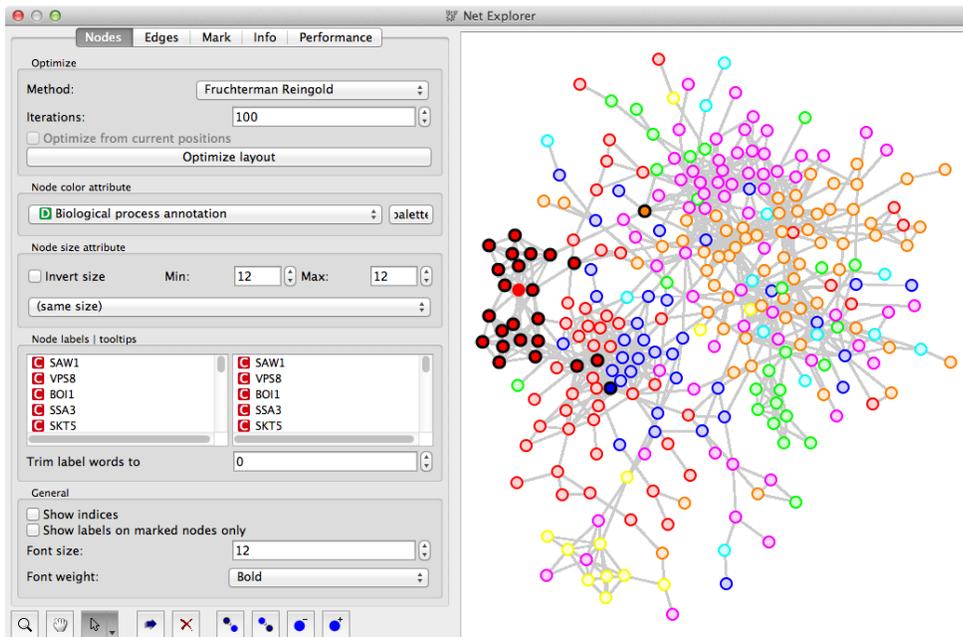
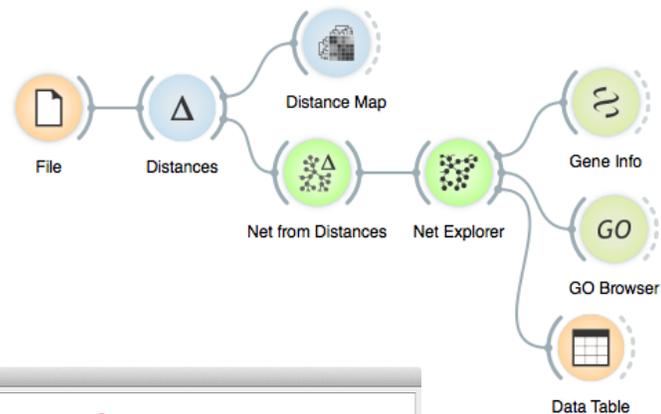
GO term	Cluster	Reference	p-value	FDR	Genes	Enrichment
proteolysis	25 (62.50%)	339 (5.31%)	2.5e-22	1.3e-20	YHR027C, Y...	[Bar chart]
proteasomal protein catabolic process	12 (30.00%)	121 (1.90%)	7.4e-12	1.3e-10	YGL048C, Y...	[Bar chart]
proteasome assembly	12 (30.00%)	28 (0.44%)	2.5e-19	7.6e-18	YGL048C, Y...	[Bar chart]

GO term	Cluster	Reference	p-value	FDR	Genes	Enrichment
ubiquitin-dependent protein catabolic process	24 (60.00%)	198 (3.10%)	2.5e-26	5.0e-24	YHR027C, Y...	[Bar chart]
proteasomal ubiquitin-independent protein catabolic process	9 (22.50%)	14 (0.22%)	3.0e-16	8.4e-15	YOR157C, Y...	[Bar chart]
proteasome assembly	12 (30.00%)	28 (0.44%)	2.5e-19	7.6e-18	YGL048C, Y...	[Bar chart]

Lesson 13: The Genetic Landscape of a Cell

The title of this lesson comes from the famous Constanzo et al. (2010) Science paper. We use a sample of their gene interaction data to reconstruct the correlation-based gene network. In this data set genes are described with their interaction profiles. We use the absolute Pearson correlation coefficient to estimate distances between genes (Distances widget). Two genes in the network are connected if their profile distance is below a certain threshold (Net from Distances). We explore the “gene galaxy” (Net Explorer) for GO function and process enrichment (GO Browser).

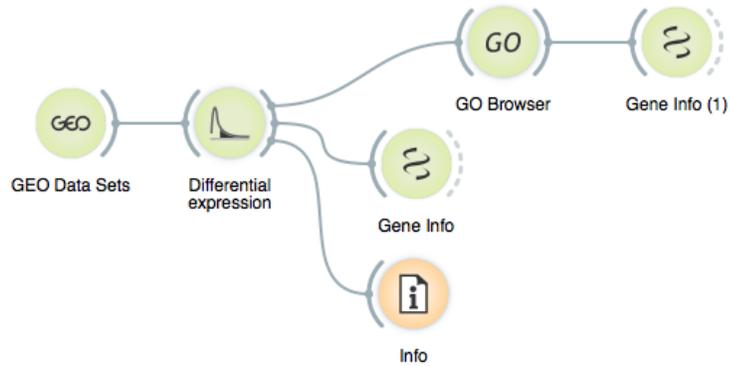
The data set for this lesson is in the documentation data sets (File widget, yeast-interactions.tab). This is a sample with 454 query genes with a subset of 184 most informative array genes selected using the CUR decomposition. Query genes in the sample were chosen to represent gene annotation groups from Figure 2 in Constanzo et al. (2010).



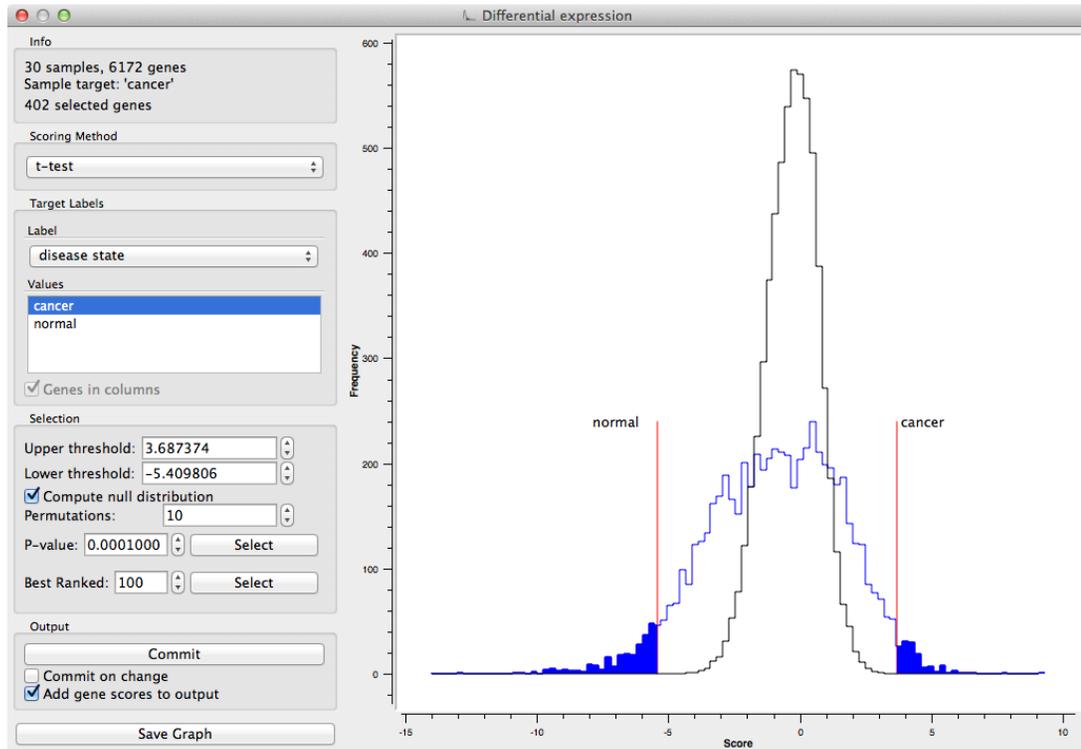
Lesson 14: Differential Expression Analysis

We can find the most differentially expressed genes in the gastric cancer data (GDS1210, 22 cases and 10 controls) with the Differential Expression widget.

Is the distribution of observed gene scores always as different from the null distribution as in GDS1210? Examine some other data sets from GEO. What can you say about those in which the observed score distribution is similar to the null distribution? Are there many such data sets in GEO?



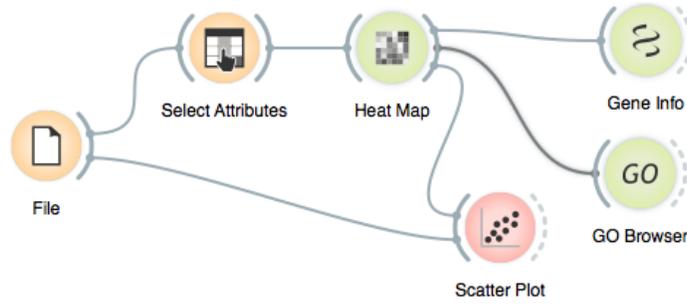
The Differential Expression widget can compare the distribution of gene scores to scores from randomly permuted data.



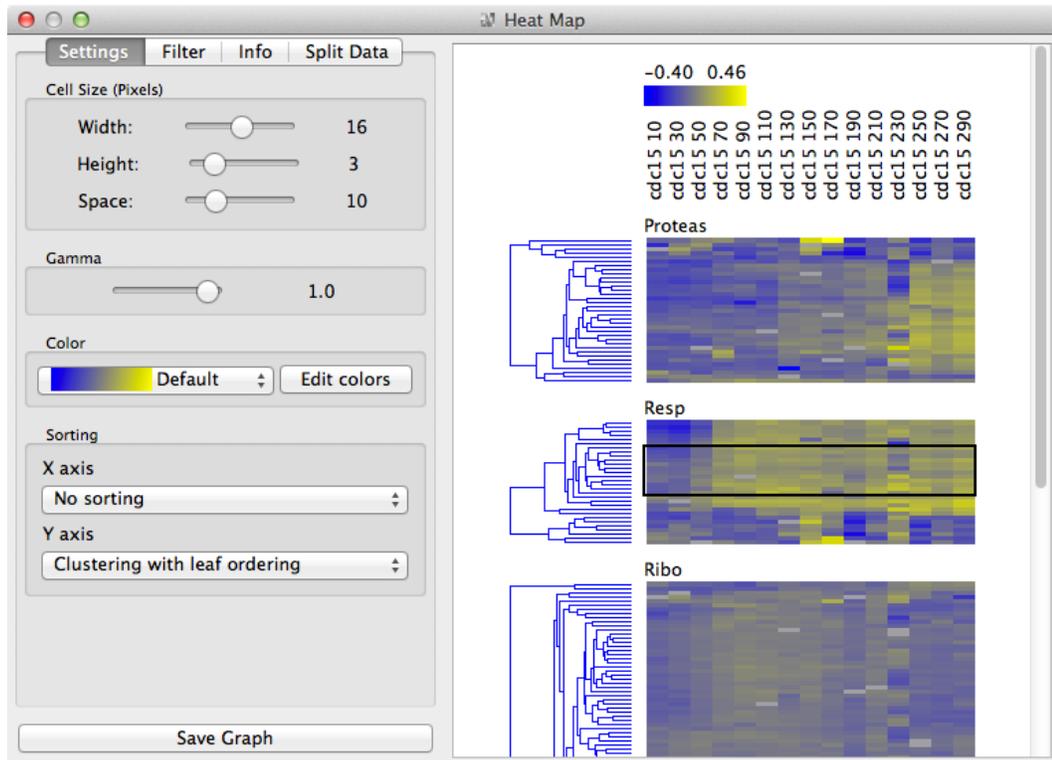
Lesson 15: Heat Maps

We can visualize gene and case profiles with a combination of a heat map and hierarchical clustering. The Heat Map widget supports row selection and outputs the associated data, which can be analyzed further (e.g. gene set enrichment analysis).

The Heat Map widget offers several ways to sort rows and columns, filter data, and define color schemes.



We use this workflow to analyze yeast cell cycle data and select a particular set of experiments using the Select Attributes widget.

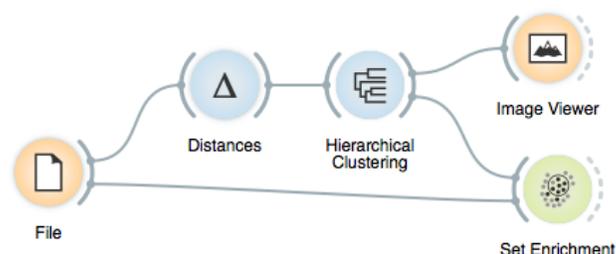


Lesson 16: Chemogenomics

A chemogenomics data set comprising 87 compounds and 289 yeast strains is sampled from Lee et al. (2014). We use data from the homozygous pool and extract compounds and strains that are found to be significant by the clearance algorithm (the `clearance_max` parameter was set to 4.00). Load the data set from the documentation data sets (File widget, `chemogenomics.tab`).

We will mine chemogenomics fitness signatures from Lee et al. (Science, 2014). In this data set compounds were characterized through fitness of yeast single-mutant strains. We will check if compounds with similar profiles share common annotations.

Orange data sets can contain links to images in local files or on the web, which can be viewed by the Image Viewer widget.



Set Enrichment

Info: 14 unique names on input

Species: None

Entity names: Entity feature: PCID

Reference: Reference set (input)

Entity Sets: MeSH, Chemicals

Commit: Commit

Report

Entities: 3, p-value: 0.0500, FDR: 0.0100

Category	Term	Count	Reference count	p-value
MeSH, Chemicals	Benzopyrans	4 (28.57%)	7 (8.05%)	0.01159
MeSH, Chemicals	Pyrans	4 (28.57%)	7 (8.05%)	0.01159
MeSH, Chemicals	Chromones	3 (21.43%)	5 (5.75%)	0.02792

Image Viewer

Info: Done: 14 images

Image Filename Attribute: Image

Title Attribute: Name

Zoom: [Slider]

Selection: Commit

Save Graph

Chemical structures shown: diclazuril, aa-861, celastrol, tpck, parthenolide, nisoldipine