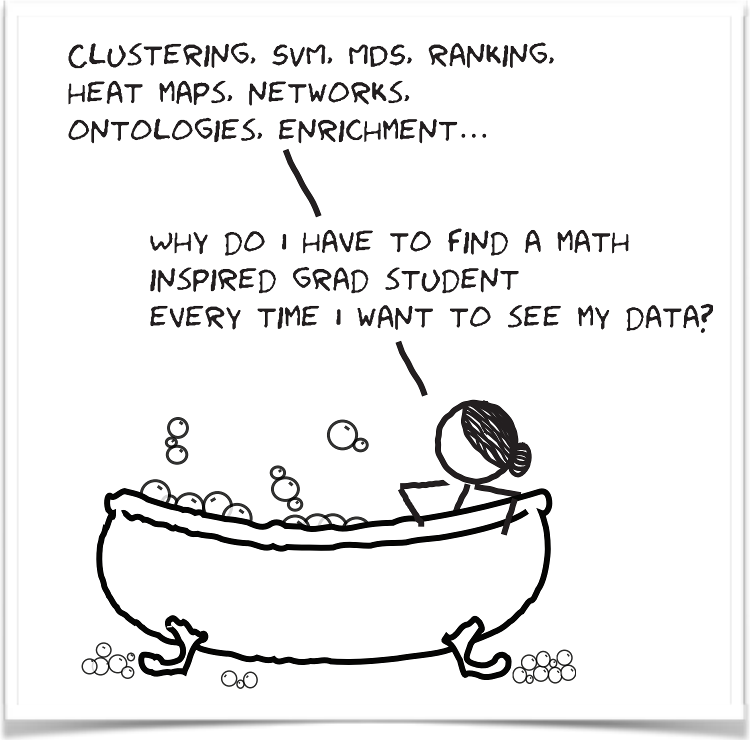

CLUSTERING, SVM, MDS, RANKING,
HEAT MAPS, NETWORKS,
ONTOLOGIES, ENRICHMENT...
WHY DO I HAVE TO FIND A MATH
INSPIRED GRAD STUDENT
EVERY TIME I WANT TO SEE MY DATA?

Data Mining w/o Programming

A Computational and Integrative Biomedical Research (CIBR) Workshop at Baylor College of Medicine

These are the working notes for the data mining workshop. The notes include Orange data mining workflows that we will construct during the workshop, and a few screenshots of interesting widgets or visualizations we will create.

Workshop instructors:
Blaz Zupan, Janez Demsar,
Marinka Zitnik, Balaji Santhanam

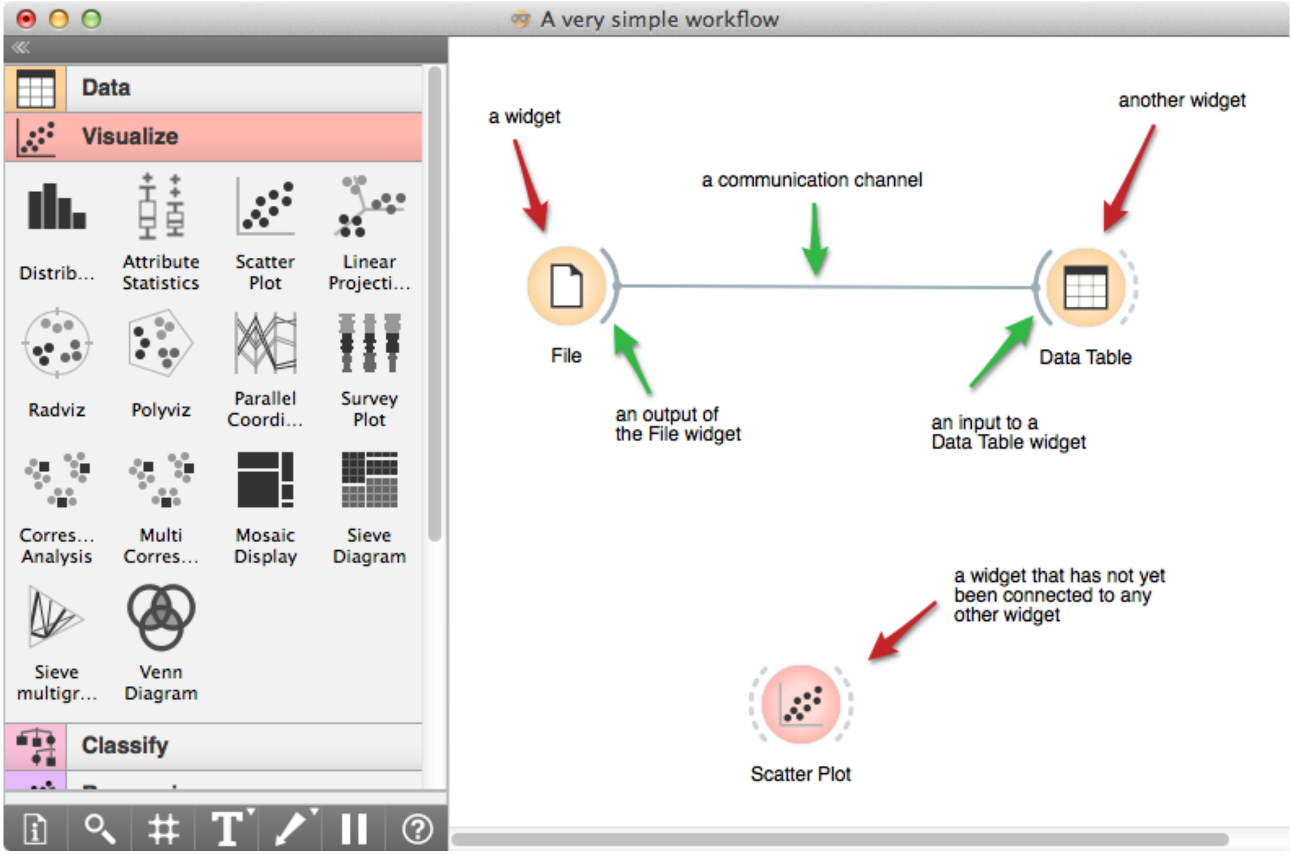
Workshop organization:
CIBR, with special thanks to
Richard Sucgang

Welcome to the hands-on Data Mining workshop! This three-hour workshop was designed for students and researchers in molecular biology to familiarize them with common data mining tasks and show how to carry them out even without programming. We will use Orange data mining suite that features visual construction of data mining flow. There are many data mining environments of this kind available today, but the organizers prefer Orange to other tools for a very simple reason - they are its authors.

This cheat sheets assumes that you installed Orange on your laptop. Otherwise, and for anybody else that will use this material later, please see the installation guide at http://biolab.github.io/bcm-workshop.

Lesson 1: Workflows in Orange

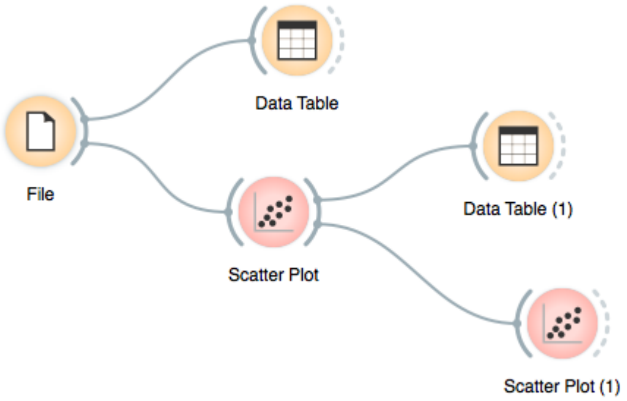
Orange workflows consists of components that read, process and visualize the data. We call them “widgets”. Widgets are chosen from a library of widgets and placed on a drawing board called “canvas”. Widgets communicate by sending tokens which are transmitted via communication channels. An output from one widget can be the input to another.



A very simple schema with two connected widgets and one “hanging” widget with no connections. Widget’s output is always on the right, and the input is on the left of the widget.

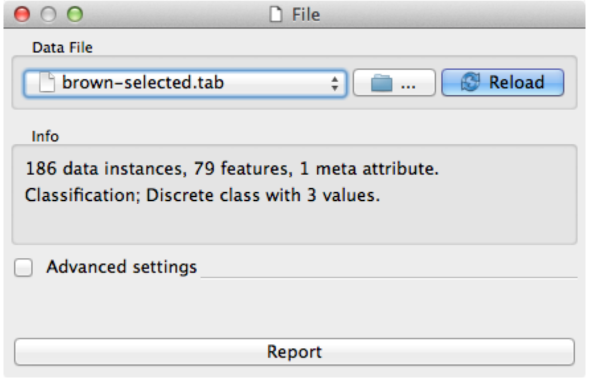
We construct workflows by placing widgets on the canvas and connecting them by dragging the line from an output “ear” of transmitting widget to an “input” ear of receiving widget.

To start, construct a data flow that consists of the File widget, two Scatter Plot widgets and a Data Table widget.



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

The File widget reads the data from the local disk. Load the data by opening the File widget. You open the widget by double clicking its icon. Orange installation comes with several preloaded data sets. From these ("Browse documentation data sets..."), choose brown-selected.tab.



Data File

brown-selected.tab

Info

186 data instances, 79 features, 1 meta attribute.
Classification; Discrete class with 3 values.

□ Advanced settings

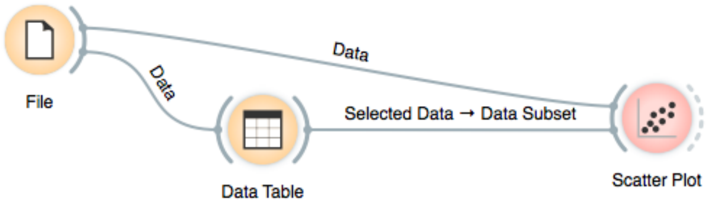
Orange workflows most often start with a File widget. The widget displays the size of the data, which for brown-selected data set include 186 rows (genes) and 81 columns. There are 79 columns with gene expression of baker's yeast under various conditions, one special "meta" column that provides for gene names and another special "class" column with gene function.

Once you load the data, open the other widgets as well. In Scatter Plot widget, select a few data points and check if they appear in Data Table (1). Use a combination of two Scatter Plots to create a "data lense", where the second Scatter Plot can show a detail of a smaller region selected in the first Scatter Plot.

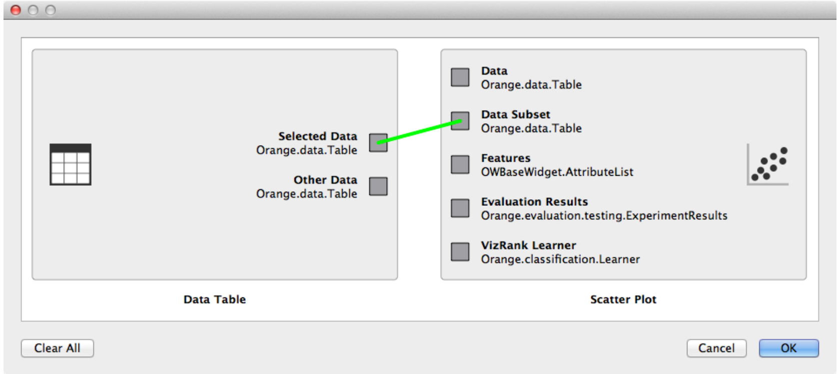
Scatter Plot for a random choice of features for this yeast expression data set does not provide much information on gene function. Does this change with a different choice of the features? Try intelligent visualization scoring by VizRank that is implemented within Scatter Plot.

We can also connect the output of the Data Table to the Scatterplot to highlight the chosen data instances (rows) in the scatterplot.

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



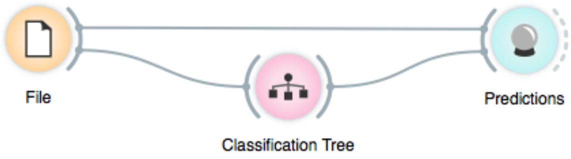
How does Orange know which is the primary data source and which one the selection? It assumes a reasonable behavior on the part of the user: the first connected signal will be the entire data set and the next one the subset. To make changes, double click on the line connecting the two widgets.



Lesson 2: Classification

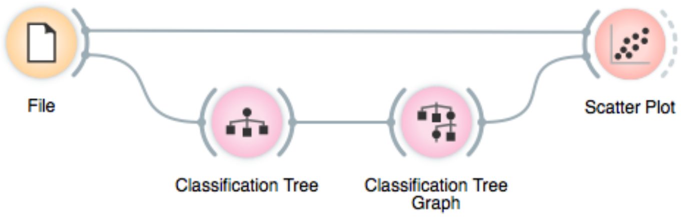
Genes in yeast data set are labeled with three functions ("Proteas", "Resp" and "Ribo"). Can we construct a model that, based on gene's gene expression profile, predicts gene function? We'll first create a model called a classification tree and observe its predictions. What's conceptually wrong here?

Something in this workflow is conceptually very wrong. What?



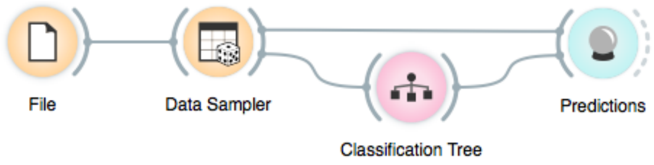
Classification trees split the data to smaller and smaller data sets where one of the data classes prevails. We can use Classification Tree Graph to visualize this model. Also, consider a combination with a Scatter Plot to demonstrate how the classification tree splits the data set to smaller and purer subsets.

Classification Tree has a data set on its input and outputs a classification tree model. This is received by a Classification Tree Graph that renders the tree. Selecting a tree node in this widget will output its corresponding data.



In the next workflow we will split the data set to two subsets: a training set and a test set. We will construct the model from the training set only, and then observe the predicted class probabilities for data instances in the test set. Are predictions reasonably correct? How can we assess their accuracy?

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.

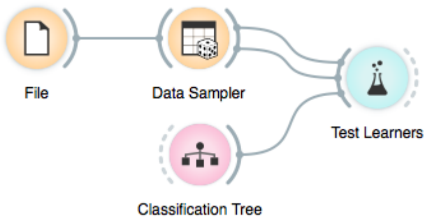


To see what the Data Sampler does, you can feed its output to Data Table or to Info widgets.

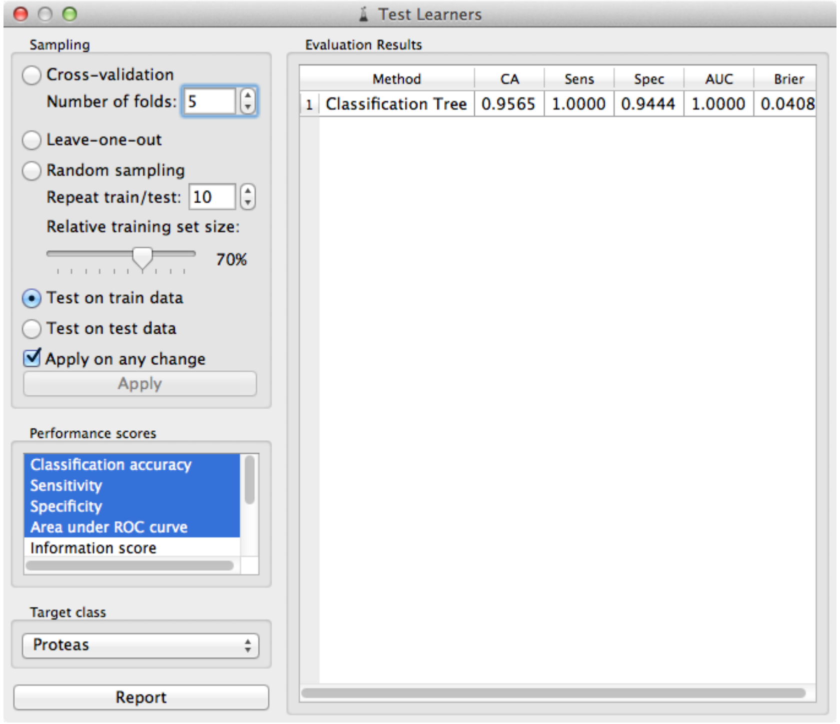
Lesson 3: Classification Accuracy

We can split the data to training and test set to measure the accuracy of the constructed model. We always need to assess the accuracy on new data, that is, on the data that has not been used for inference of prediction model. Accuracy can then be measured as proportion of data instances in the test set for which the class prediction was correct.

Try changing the size of the training data set and observe the impact on the accuracy. What do you expect? Try this with some other classification data sets, say from Orange's documentation data sets.



Classification accuracy depends on prior class distribution in the training set. Consider a data set with 90% of instances labeled with the same class, and a predictor with 70% accuracy. Evaluation scores such as Area Under ROC are not affected by this and can hence serve as a better scoring function.



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

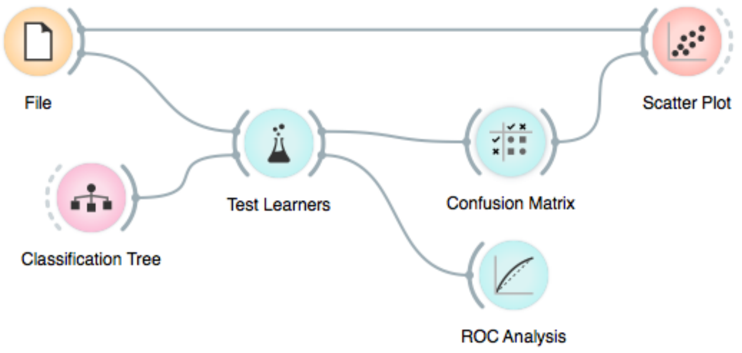
Evaluation Results

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

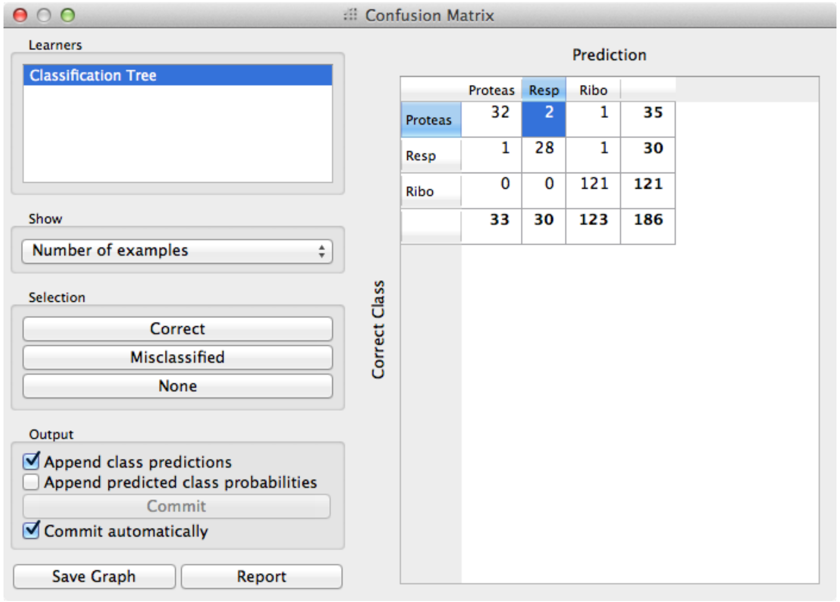
Lesson 4: Cross-Validation

Accuracy scoring may depend on particular split of the data to a training and test set. Instead of measuring the accuracy just once, we can repeat the measurement several times, each time choosing a different data sample for training. One of such schemes called cross-validation can be invoked by Test Learners widget. We will analyze its output by confusion matrix and ROC curve.

Selection of cells in Confusion Matrix place their related data instances to the output of the widget. In this schema we send them to Scatter Plot to be visualized as a data subset. What can you say about misclassified data? Does scatter plot provide any insights? Outliers?



A feature of cross-validation is that each data instance is used for testing exactly once. We can then use Confusion Matrix widget to find how many test data instances were classified correctly and if not, for which class were they mistaken.



Learners: Classification Tree

Show: Number of examples

Output

☑ Append class predictions
☐ Append predicted class probabilities
☑ Commit automatically

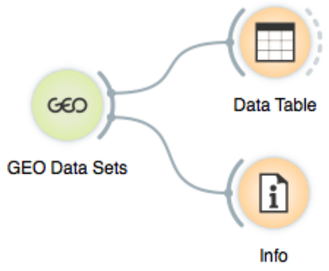
Prediction

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

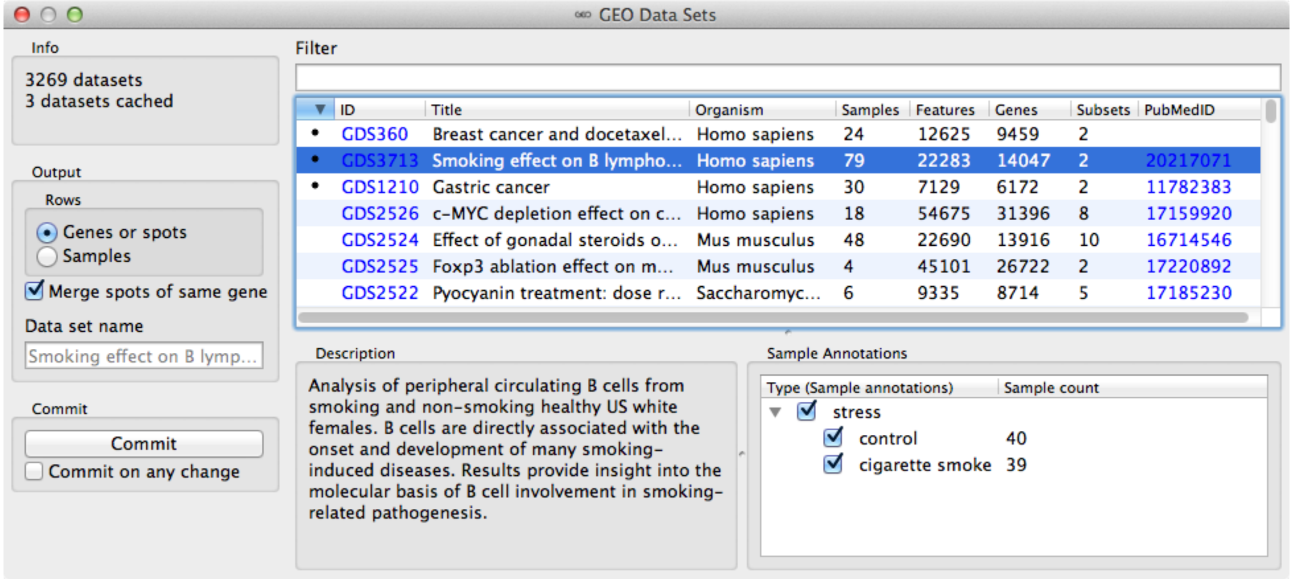
Lesson 5: GEO Data Sets

It's time to explore some other data sets. In its bioinformatics add-on, Orange provides access to a data set library by Gene Expression Omnibus (GEO). Orange queries GEO for each data set you select, and downloads it to your local disk. For a start, construct a simple workflow and load few data sets, or better, for the workshop, inspect the preloaded data sets.

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



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



GEO Data Sets

Info

3269 datasets
3 datasets cached

Output

Rows

● Genes or spots
○ Samples

☑ Merge spots of same gene

Data set name

Smoking effect on B lymp...

Commit

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.

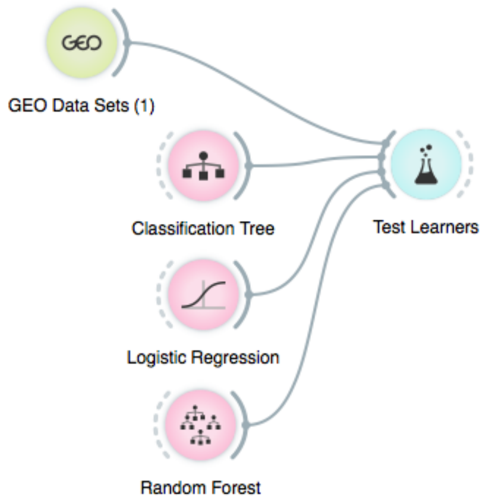
Sample Annotations

Type (Sample annotations)	Sample count
▼ ☑ stress	
☑ control	40
☑ cigarette smoke	39

Lesson 6: GEO Data Sets and Classification

From 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?

Random Forest often achieves very good accuracy on gene expression data sets. Try changing the number of classification trees included in the forest. How does the accuracy change? Do forests beat a single classification trees? Why? How does logistic regression compare to the other two methods?

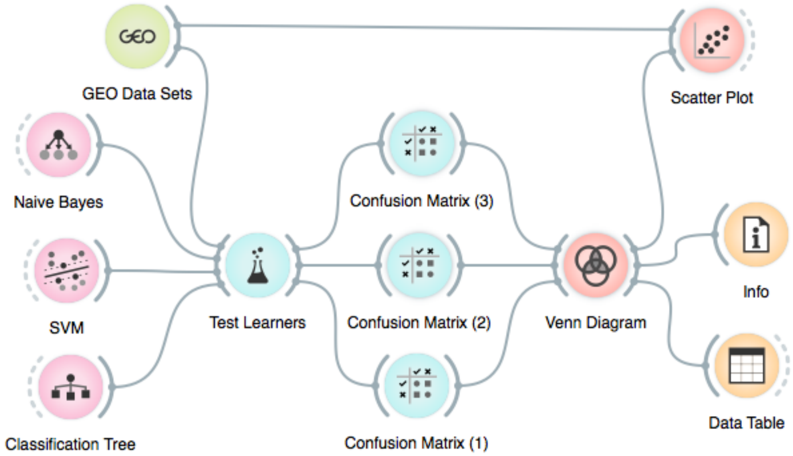


Besides classification trees, we here also included logistic regression and random forest. Some of these algorithms take time, so for a start better choose data sets with fewer cases and features.

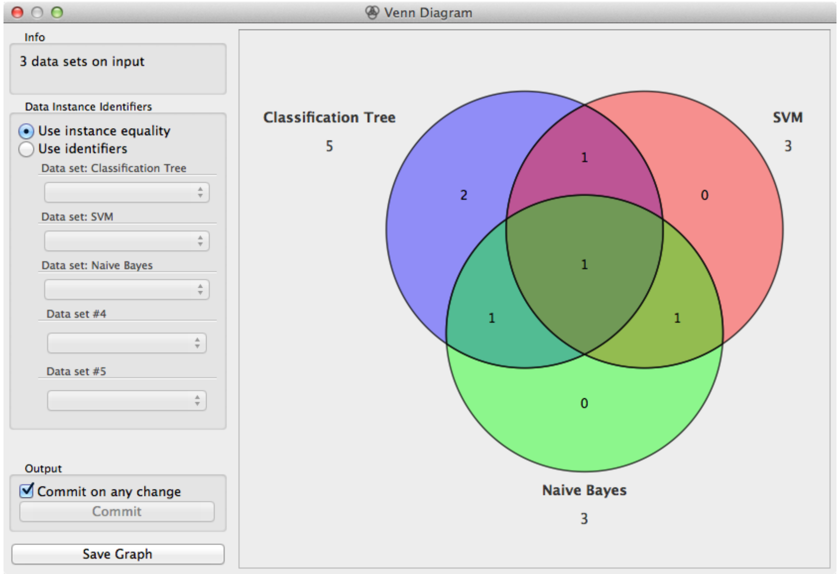
Lesson 7: Venn Diagram

We here admit that the following schema looks a bit crazy. But it's not as complicated as it looks. The question we had was: do different classifiers misclassify the same tissue samples? That is, are there some cases that are really hard to classify? Could they be outliers, or even originally misclassified tissue samples? While we can't computationally answer the last question, we can provide the answers to all other questions by cross-validating the classifiers, selecting misclassified instances in the Confusion Matrix, and then relating the three sets of misclassifications in the Venn diagram.

Most of the widgets in Orange are interactive. For example, you can click on different sections of 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.



Info

3 data sets on input

Data Instance Identifiers

● Use instance equality
○ Use identifiers

Data set: Classification Tree

Data set: SVM

Data set: Naive Bayes

Data set #4

Data set #5

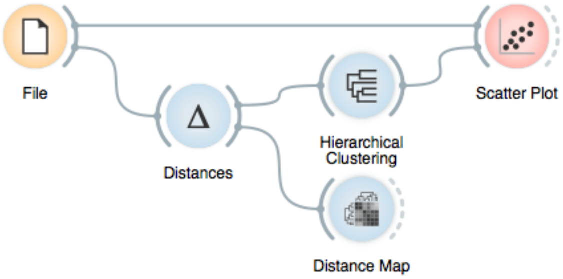
Output

☑ Commit on any change

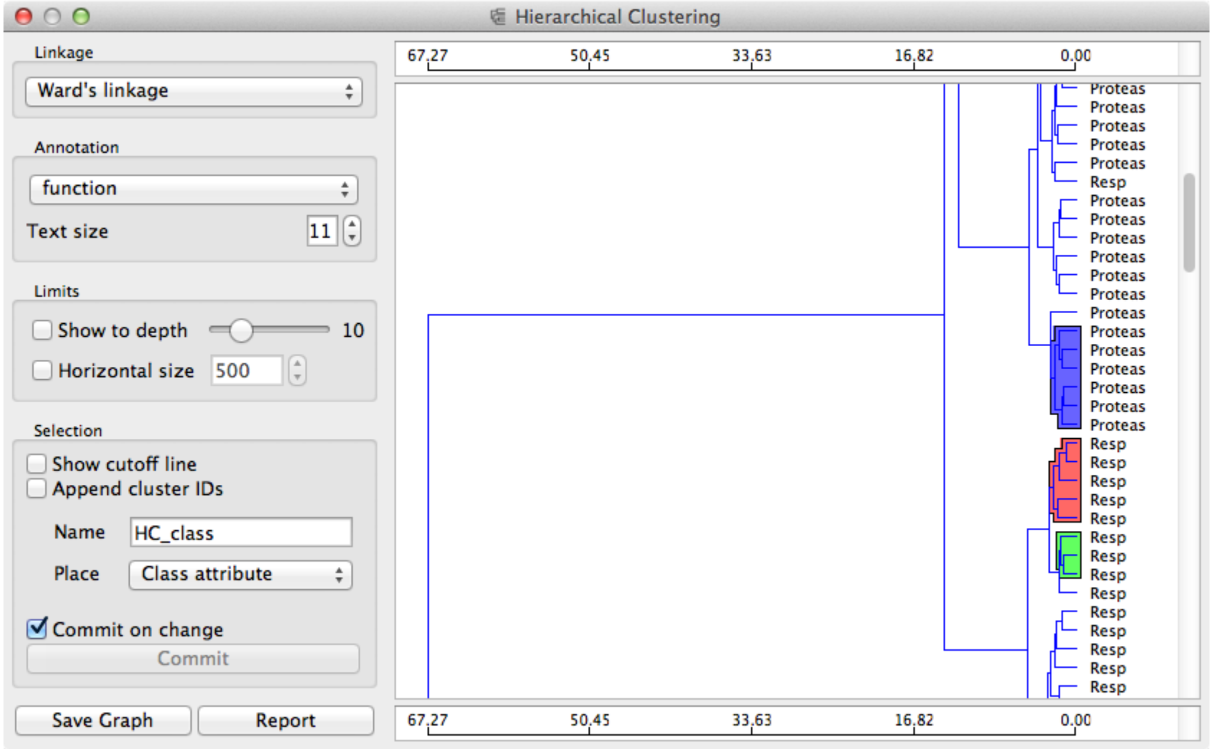
Lesson 8: Hierarchical Clustering

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

We have used Euclidean distance (in Distances) and Ward's linkage (in Hierarchical Clustering). Euclidean distance may not be the best choice. Why? Try some other distances instead. Any changes in the dendrogram?



We have decided to display the data instances selected in the dendrogram as a subset in a Scatter Plot. Make sure this widget is showing and informative visualization. (What would that be?)


Hierarchical Clustering

Linkage
Ward's linkage

Annotation
function
Text size 11

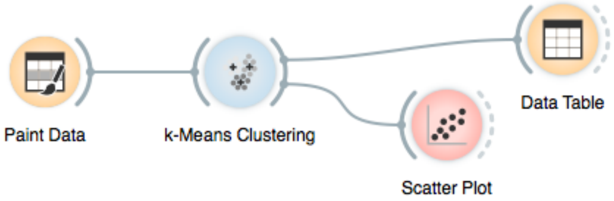
Limits
☐ Show to depth 10
☐ Horizontal size 500

Selection
☐ Show cutoff line
☐ Append cluster IDs
Name HC_class
Place Class attribute

☑ Commit on change

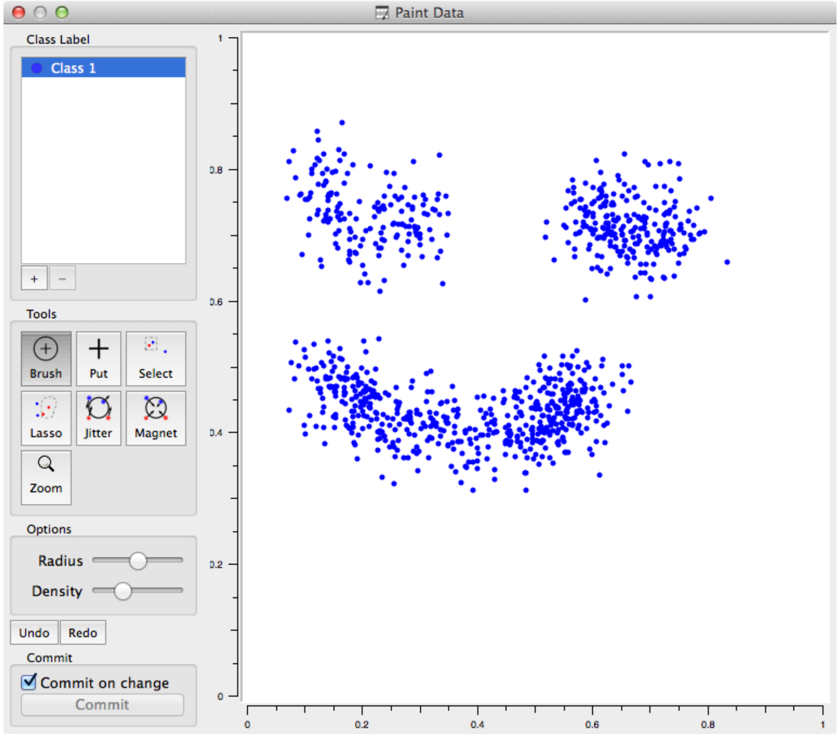
Lesson 9: k - Means Clustering

Hierarchical clustering fails on larger data sets due to prohibitive size of a distance matrix. An alternative approach is conglomerative clustering. Here we have to provide the number of clusters in advance, or a range for this parameter and a cluster scoring technique will find the optimal number of clusters. While you are free to use conglomerative clustering with any of the data sets we have examined so far (you *should* try this, actually), we really like to discuss the properties of this method with hand-painted data sets.



The game we like to play is to see if silhouette scoring in k-means can figure out "the right" number of clusters.

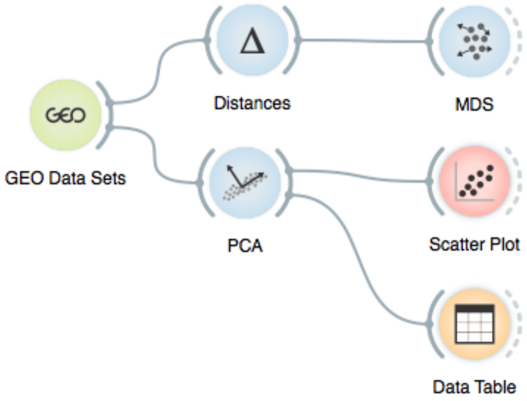
How many clusters should be identified for the data set on the right? Why? What is the number of clusters proposed by the silhouette method and k-means clustering? Could you "correct" this data set to help k-means find the expected number of clusters?



Lesson 10: Data Projection

We have already worked with one data projection - a scatter plot. But this one only shows the data projected onto a hyperplane defined by two data dimensions. A technique to find projections that expose the largest variance within the data is called Principal Component Analysis (PCA). A different approach is Multidimensional Scaling (MDS), where data is projected to Euclidean space such that the distances between points in the visualization that matches the distances between the data points. Note that MDS requires distances. The two approaches often yield very similar visualizations.

Instead of the GEO Data Set experiment also with brown-selected.tab (File widget). How different are the visualizations by PCA and MDS? Which one better recapitulates the data?

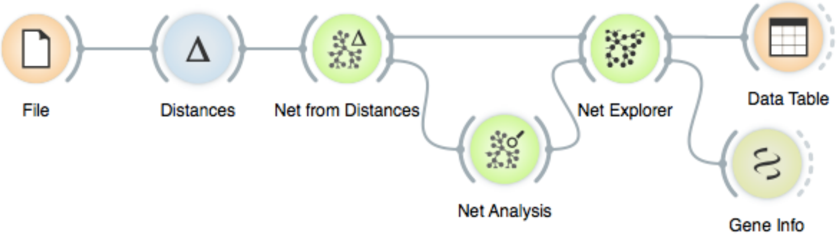


PCA can be also used for preprocessing, outputting the data in a new feature space to be then used for, say, classification. This could sometimes help to increase accuracy, but also makes the results very harder to interpret (why?).

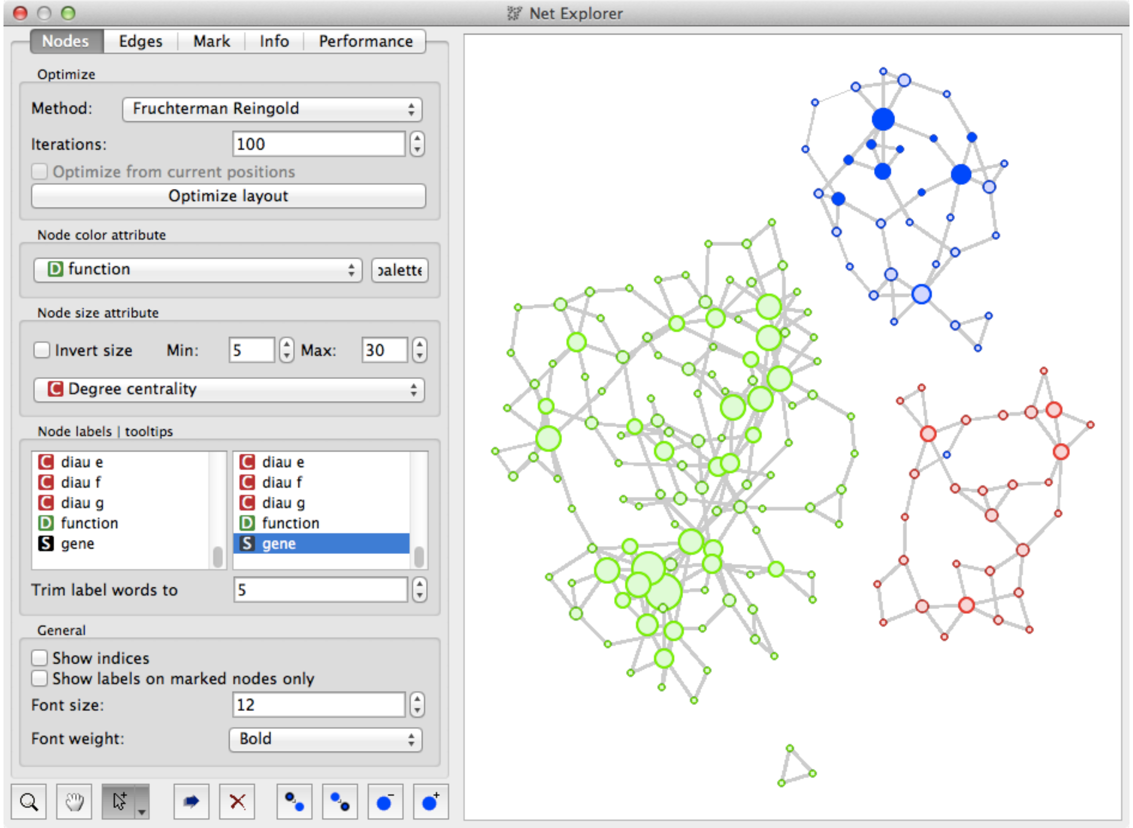
Lesson 11: Correlation Networks

Similarity between data instances (genes, tissue samples, chemicals) may also be presented with the networks. We need to decide on similarity threshold, or limit the number of edges per node, or both. You need to have Orange network add-on 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 choices of distance metrics? What are the hub genes?



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



Optimize

Method: Fruchterman Reingold

Iterations: 100

□ Optimize from current positions

Optimize layout

Node color attribute

D function ɔalette

Node size attribute

□ Invert size Min: 5 Max: 30

C Degree centrality

Node labels | tooltips

C diau e	C diau e
C diau f	C diau f
C diau g	C diau g
D function	D function
S gene	S gene

Trim label words to 5

General

□ Show indices

□ Show labels on marked nodes only

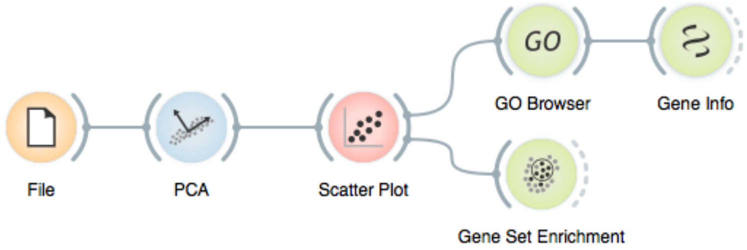
Font size: 12

Font weight: Bold

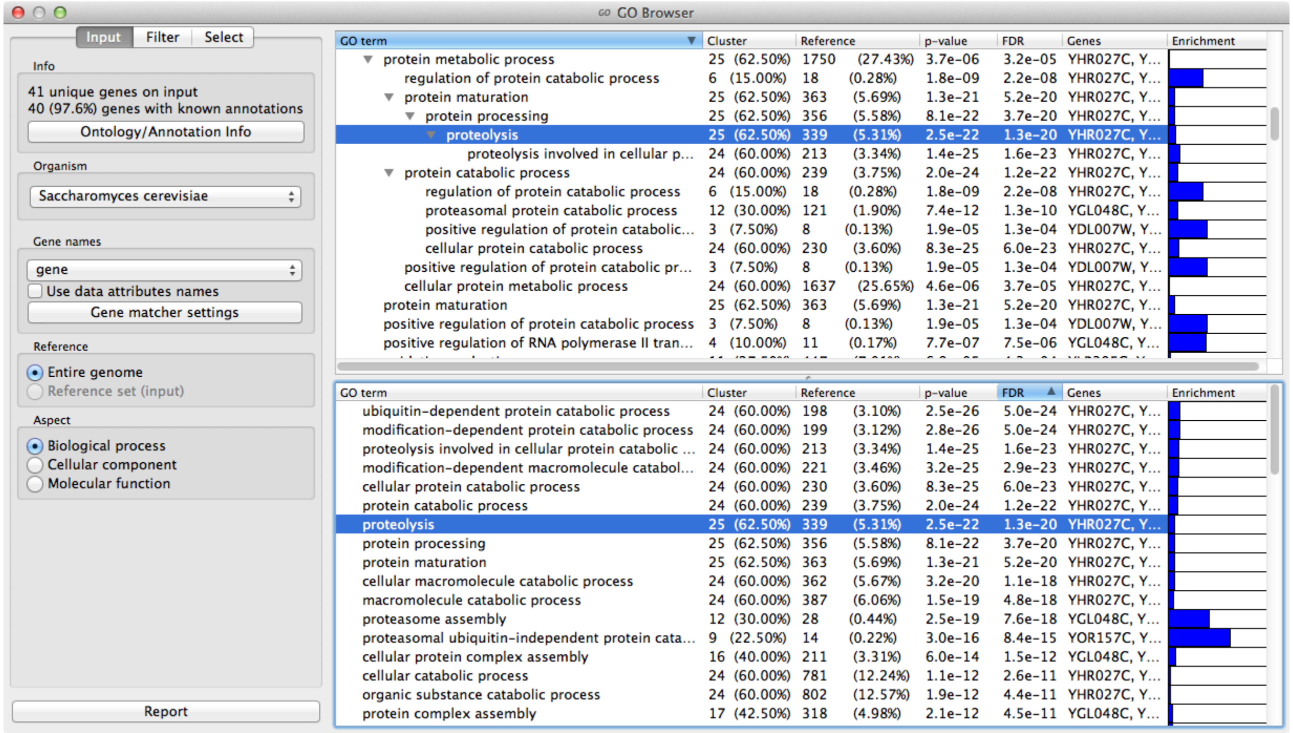
Lesson 12: Gene Set Enrichment

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

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



GO Browsers presents two views of enriched pathways, one reporting on the ontology tree and the other presenting a list of enriched GO terms.



Info

41 unique genes on input
40 (97.6%) genes with known annotations

Ontology/Annotation Info

Organism: Saccharomyces cerevisiae

Gene names: gene

□ Use data attributes names

Gene matcher settings

Reference

● Entire genome
○ Reference set (input)

Aspect

● Biological process
○ Cellular component
○ Molecular function

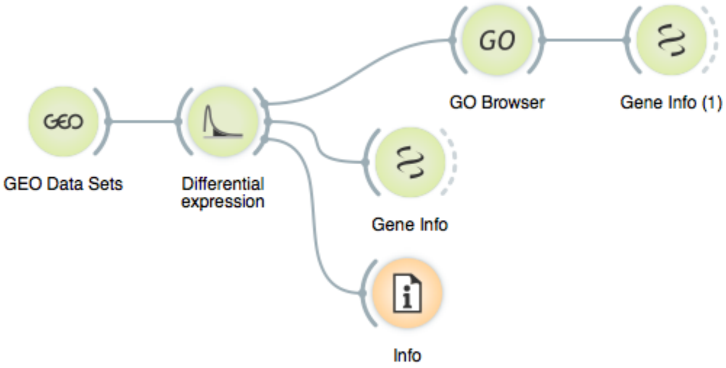
GO term	Cluster	Reference	p-value	FDR	Genes	Enrichment
▼ protein metabolic process	25 (62.50%)	1750 (27.43%)	3.7e-06	3.2e-05	YHR027C, Y...	
regulation of protein catabolic process	6 (15.00%)	18 (0.28%)	1.8e-09	2.2e-08	YHR027C, Y...	
▼ protein maturation	25 (62.50%)	363 (5.69%)	1.3e-21	5.2e-20	YHR027C, Y...	
▼ protein processing	25 (62.50%)	356 (5.58%)	8.1e-22	3.7e-20	YHR027C, Y...	
▼ proteolysis	25 (62.50%)	339 (5.31%)	2.5e-22	1.3e-20	YHR027C, Y...	
proteolysis involved in cellular p...	24 (60.00%)	213 (3.34%)	1.4e-25	1.6e-23	YHR027C, Y...	
▼ protein catabolic process	24 (60.00%)	239 (3.75%)	2.0e-24	1.2e-22	YHR027C, Y...	
regulation of protein catabolic process	6 (15.00%)	18 (0.28%)	1.8e-09	2.2e-08	YHR027C, Y...	
proteasomal protein catabolic process	12 (30.00%)	121 (1.90%)	7.4e-12	1.3e-10	YGL048C, Y...	
positive regulation of protein catabolic...	3 (7.50%)	8 (0.13%)	1.9e-05	1.3e-04	YDL007W, Y...	
cellular protein catabolic process	24 (60.00%)	230 (3.60%)	8.3e-25	6.0e-23	YHR027C, Y...	
positive regulation of protein catabolic pr...	3 (7.50%)	8 (0.13%)	1.9e-05	1.3e-04	YDL007W, Y...	
cellular protein metabolic process	24 (60.00%)	1637 (25.65%)	4.6e-06	3.7e-05	YHR027C, Y...	
protein maturation	25 (62.50%)	363 (5.69%)	1.3e-21	5.2e-20	YHR027C, Y...	
positive regulation of protein catabolic process	3 (7.50%)	8 (0.13%)	1.9e-05	1.3e-04	YDL007W, Y...	
positive regulation of RNA polymerase II tran...	4 (10.00%)	11 (0.17%)	7.7e-07	7.5e-06	YGL048C, Y...	

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...	
modification-dependent protein catabolic process	24 (60.00%)	199 (3.12%)	2.8e-26	5.0e-24	YHR027C, Y...	
proteolysis involved in cellular protein catabolic ...	24 (60.00%)	213 (3.34%)	1.4e-25	1.6e-23	YHR027C, Y...	
modification-dependent macromolecule catabol...	24 (60.00%)	221 (3.46%)	3.2e-25	2.9e-23	YHR027C, Y...	
cellular protein catabolic process	24 (60.00%)	230 (3.60%)	8.3e-25	6.0e-23	YHR027C, Y...	
protein catabolic process	24 (60.00%)	239 (3.75%)	2.0e-24	1.2e-22	YHR027C, Y...	
proteolysis	25 (62.50%)	339 (5.31%)	2.5e-22	1.3e-20	YHR027C, Y...	
protein processing	25 (62.50%)	356 (5.58%)	8.1e-22	3.7e-20	YHR027C, Y...	
protein maturation	25 (62.50%)	363 (5.69%)	1.3e-21	5.2e-20	YHR027C, Y...	
cellular macromolecule catabolic process	24 (60.00%)	362 (5.67%)	3.2e-20	1.1e-18	YHR027C, Y...	
macromolecule catabolic process	24 (60.00%)	387 (6.06%)	1.5e-19	4.8e-18	YHR027C, Y...	
proteasome assembly	12 (30.00%)	28 (0.44%)	2.5e-19	7.6e-18	YGL048C, Y...	
proteasomal ubiquitin-independent protein cata...	9 (22.50%)	14 (0.22%)	3.0e-16	8.4e-15	YOR157C, Y...	
cellular protein complex assembly	16 (40.00%)	211 (3.31%)	6.0e-14	1.5e-12	YGL048C, Y...	
cellular catabolic process	24 (60.00%)	781 (12.24%)	1.1e-12	2.6e-11	YHR027C, Y...	
organic substance catabolic process	24 (60.00%)	802 (12.57%)	1.9e-12	4.4e-11	YHR027C, Y...	
protein complex assembly	17 (42.50%)	318 (4.98%)	2.1e-12	4.5e-11	YGL048C, Y...	

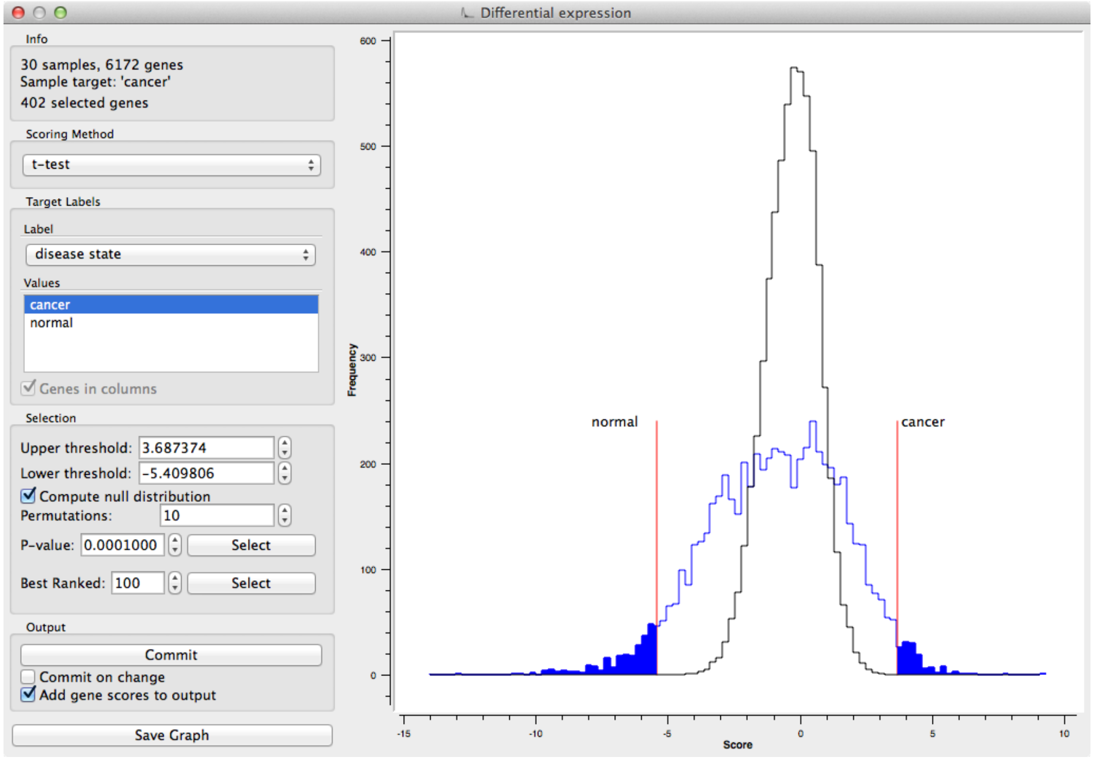
Lesson 13: Differential Expression Analysis

Using, say, data on gastric cancer (GDS1210, 22 cases and 10 controls), we can find the most differentially expressed genes using Differential Expression widget.

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



Info

30 samples, 6172 genes
Sample target: 'cancer'
402 selected genes

Scoring Method: t-test

Target Labels

Label: disease state

Values: cancer, normal

☑ Genes in columns

Selection

Upper threshold: 3.687374
Lower threshold: -5.409806
☑ Compute null distribution
Permutations: 10
P-value: 0.0001000
Best Ranked: 100

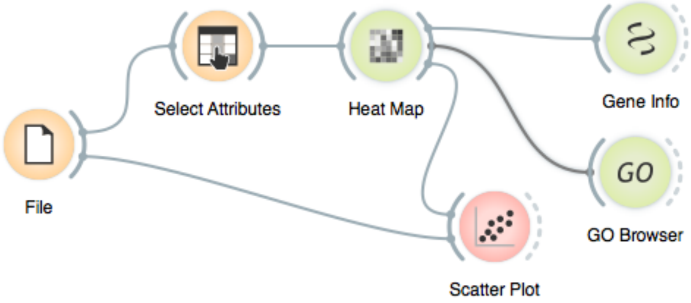
Output

☐ Commit on change
☑ Add gene scores to output

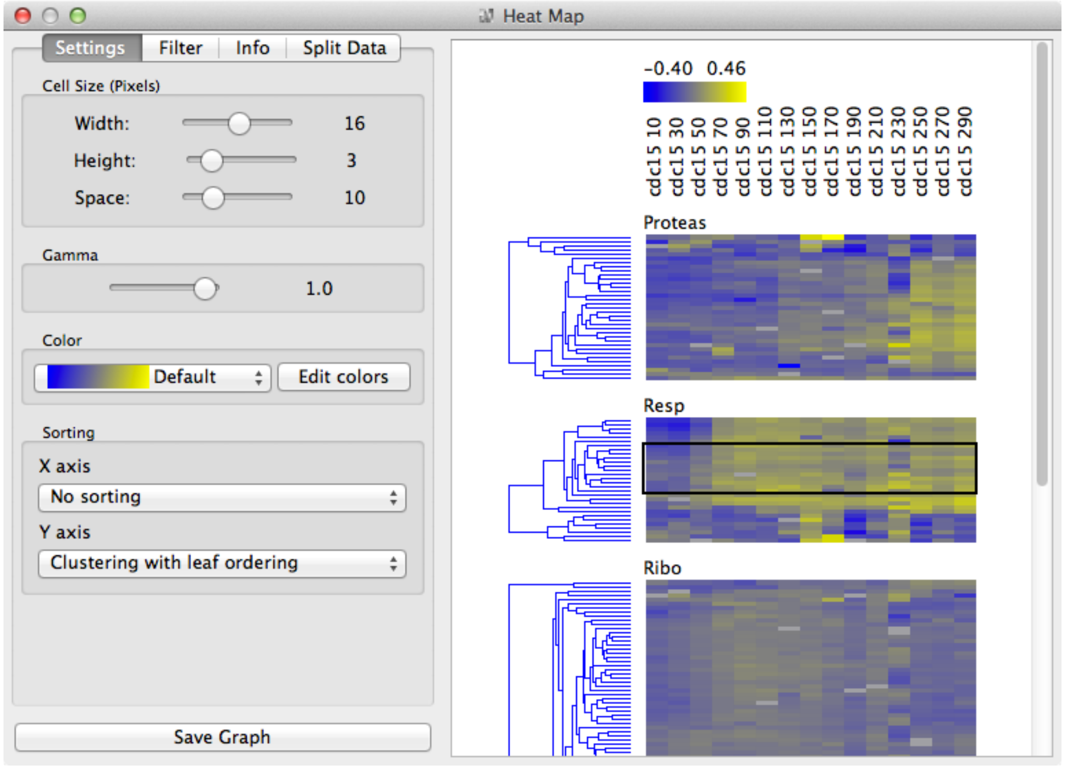
Lesson 14: Heatmaps

Heatmap combined with hierarchical clustering is a cool way to visualize gene and case profiles. Orange's Heatmap widget supports row selection and outputs the data that can be analyzed further, say for gene set enrichment analysis.

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



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


Heat Map

Settings | Filter | Info | Split Data

Cell Size (Pixels)

Width: 16

Height: 3

Space: 10

Gamma

1.0

Color

Default | Edit colors

Sorting

X axis

No sorting

Y axis

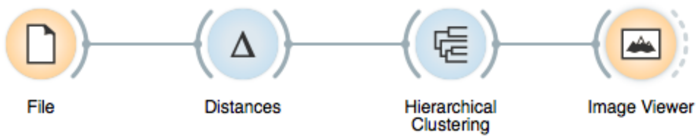
Clustering with leaf ordering

Save Graph

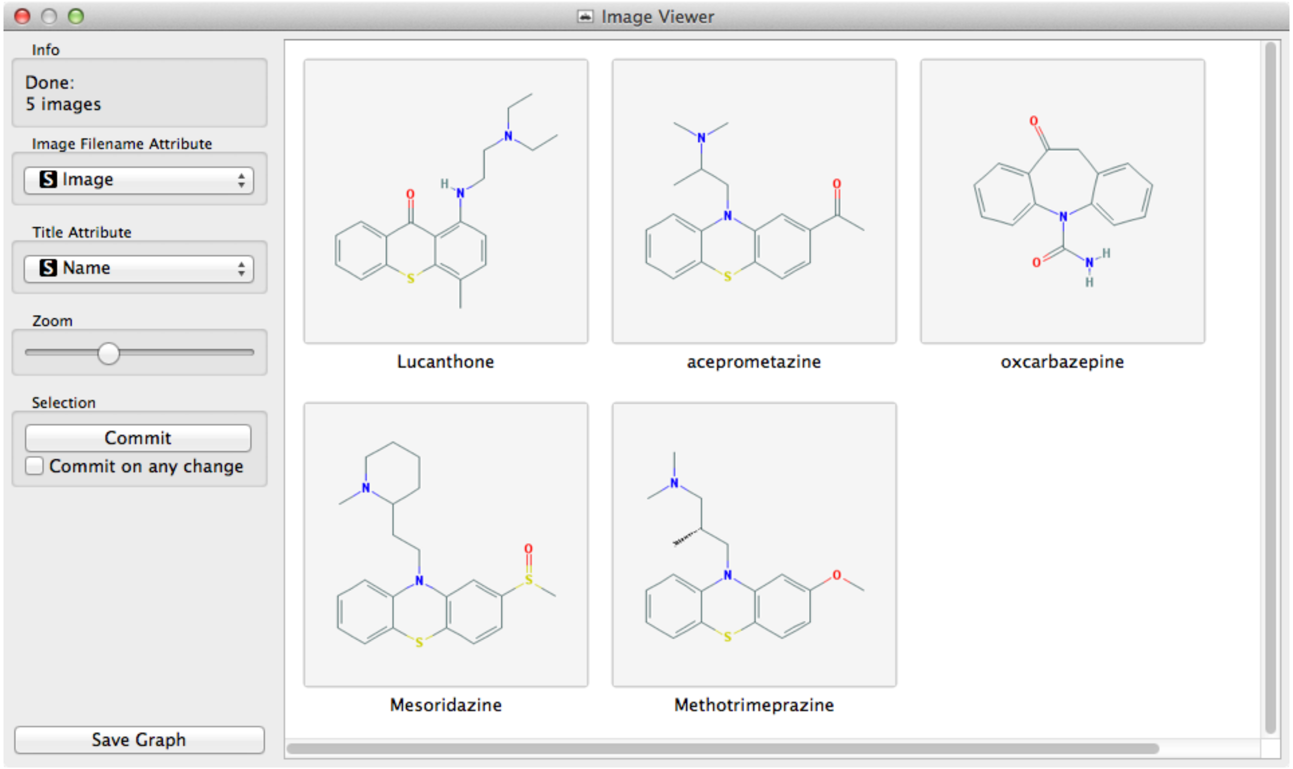
Lesson 15: Working with Images

Orange data sets can contain links to images in local files or on the web. Image Viewer widget displays images that are included in the input data set.

Image links are included in the data as meta attributes, that is, columns that are otherwise excluded from any numerical analysis. Check this out by inspecting the data in a Data Table widget. What is the type of the column that holds chemical names?



We use this workflow on a subset of Drug bank's chemicals with pharmacological actions and fingerprints to characterize chemical structures. Load this data set from http://goo.gl/NZVrZ7.



Done:
5 images

Image Filename Attribute: Image

Title Attribute: Name

□ Commit on any change