ARco

a Bitlab software

Association Rules collaborative tool

Integrated suite for association rule discovering in medical and molecular data

---

User Manual

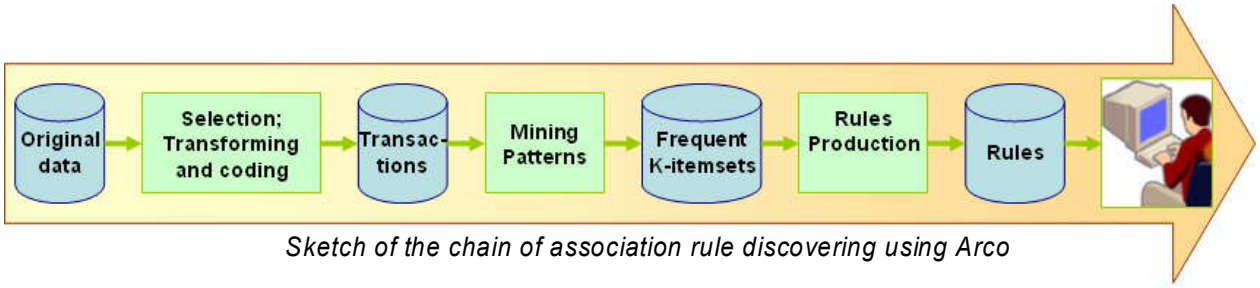
Version v1: 8th November 2007.
On-line updated information available at:
http://chirimoyo.ac.uma.es/arco

Developed by:
Jesús Jiménez Espada
Javier Rios
Andrés Rodríguez
Oswaldo Trelles

Report incidences to:
ots@ac.uma.es

ARco pipeline

As described in the Introduction section, Arco has been organised to fulfil the KDD procedure integrating a diverse gallery of methods with different but combined scope. At the end, or as one important part of KDD, we devise **ARco** that should take place in the data selection, transforming, processing and high level analysis, including visualization (for human analysis) of the new expressed knowledge in the form of association rules or the co-occurrence of events from which is possible to produce a conclusion with certain degree of confidence. Next picture depict a sketch of this chain as we see it.



Sketch of the chain of association rule discovering using Arco

First step in this chain is the selection of the data relevant to be subject of analysis. Over this selected dataset is necessary, in general, operate on it to focus the processes in particular features. Data transformations, reduction and compacting, hierarchical simplification, diverse alternative coding procedures, etc. are important procedures in this step. A collection of transactions in the form of a list of numbers that represent events that co-occur simultaneously is the resulting output. This output is the input to identify k-itemset (set of "k" items appearing together more frequently than expected by chance). From these frequent k-itemsets it is possible infer rules with certain confidence (estimated from the dataset).

Steps

ARco is endowed with different algorithms to be applied on the same data set in pipeline fashion. This guided tour will shown each of these procedures, in the following order:

a) Installation guide
b) Load Step which includes filtering and transforming data to produce a transaction dataset
c) Mining transactions to identify frequent k-itemsets
d) Ruling the frequent k-itemset to produce rules
e) Analysis procedures.

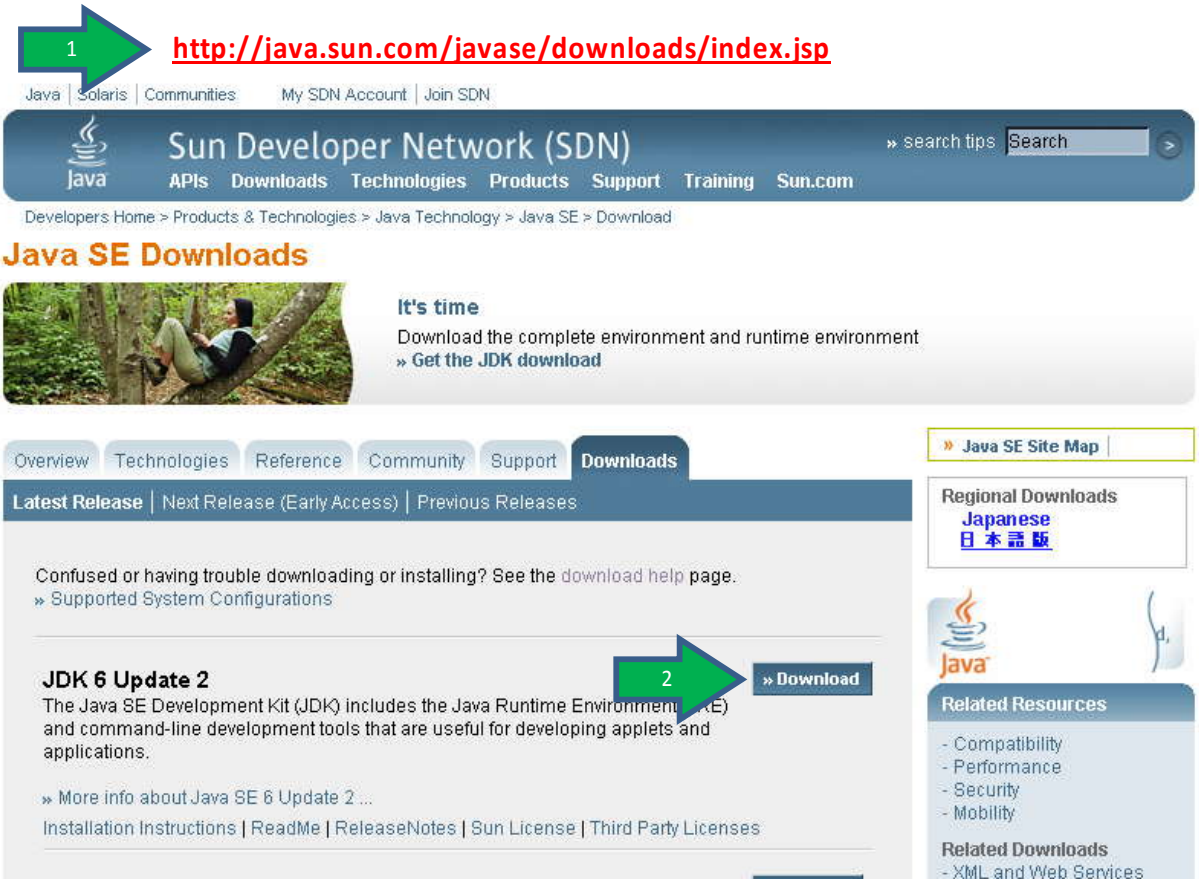
ARco installation guide

Java support has been chosen with the aim to extend the scope of ARco. Installing a Java virtual machine available for most of the current operating systems is enough to have a full operative environment

System Requirements

- Java virtual machine 1.50 or latter
- Last version of ARco software

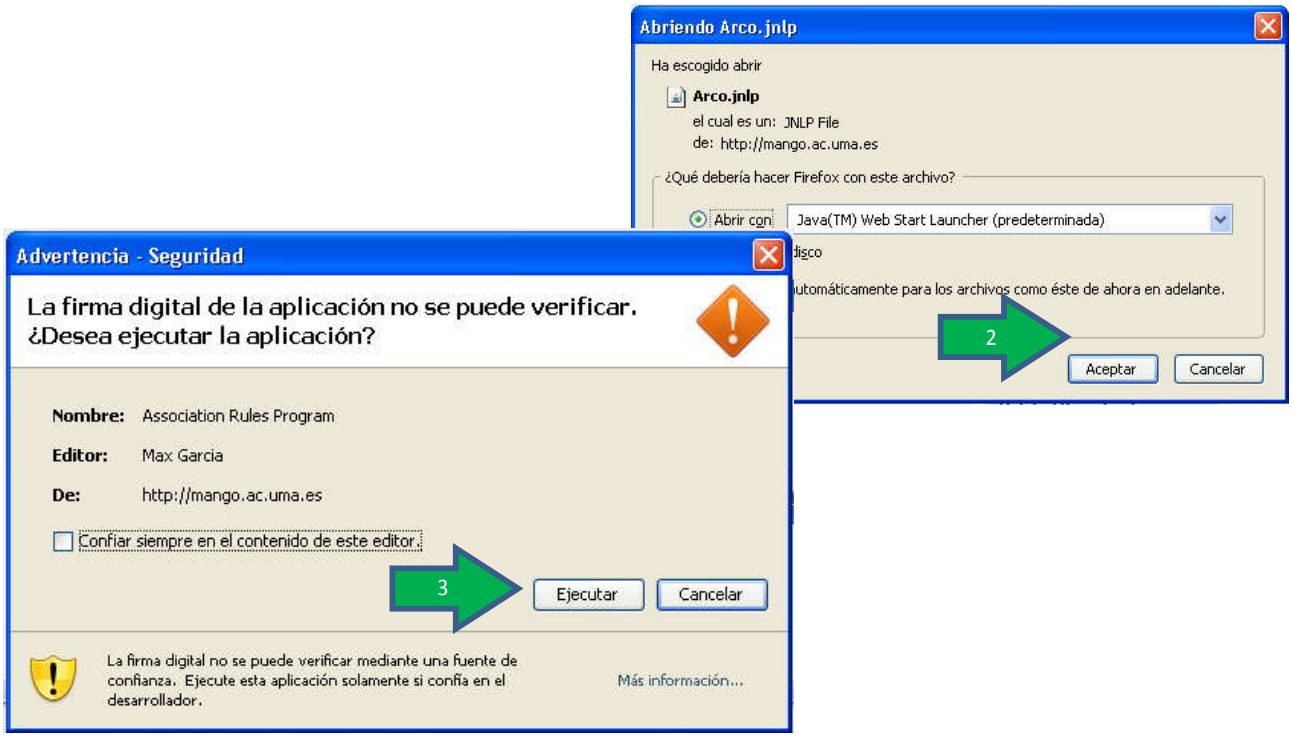
Java virtual machine



Download ARco

from http://chirimoyo.ac.uma.es/arco

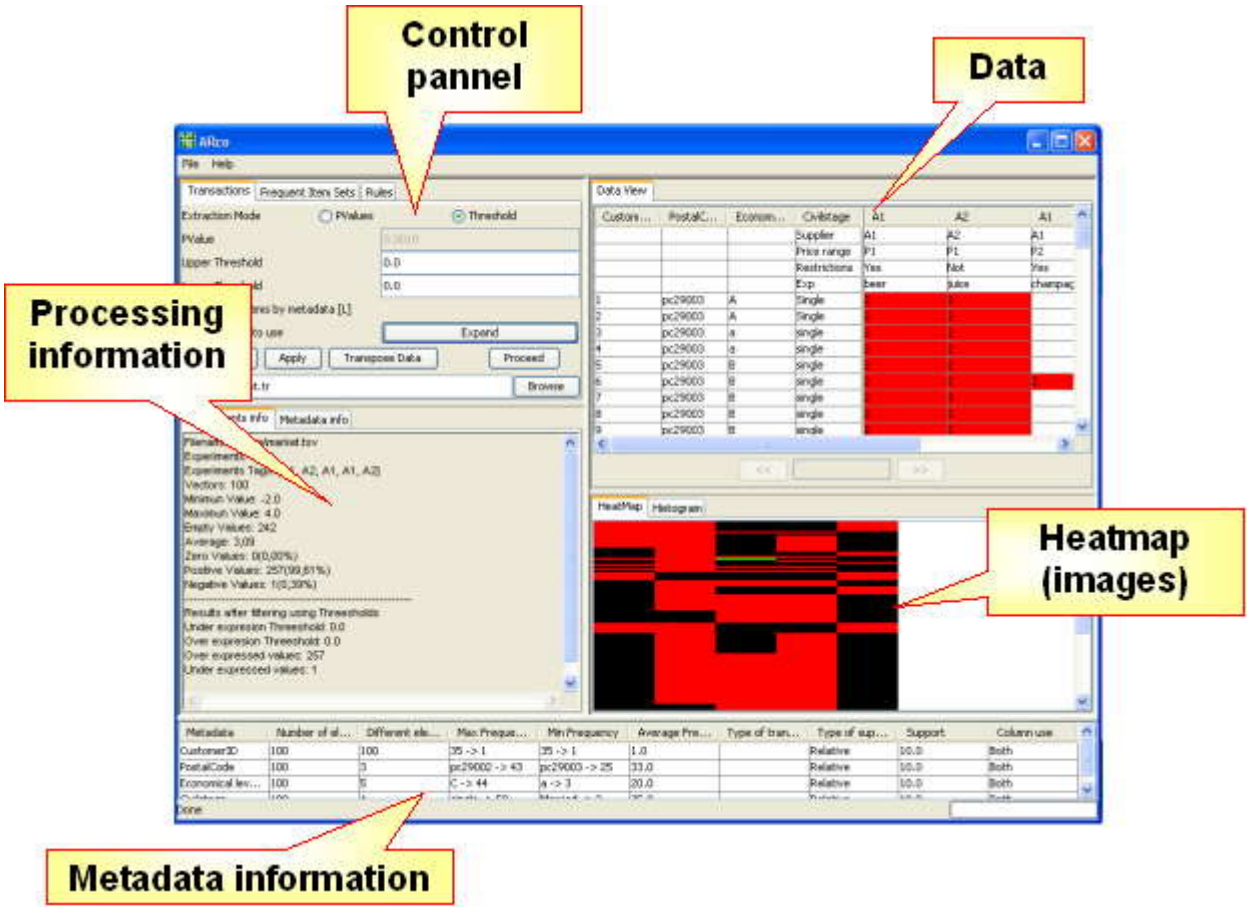
1 http://chirimoyo.ac.uma.es/arco
mango.ac.uma.es/ACGT/jaws/apps/test/Arco.jnlp



Note: Since ARco manuscript is in the evaluation process, the software is only available upon request.

ARco main screen

ARco is organised in five frames, each one with the ability to contain several sub-tabs



The most important is the “Control Pane” in which the main ARco options are available and parameters are settled. The “Data” frame contains original and processed datasets (i.e. gene-expression matrix or association rules). Below the control pane, one frame is devoted to display summarised information about data processing; and also have a tab for graphical displaying of rules. Heatmap frame contains different data representations and on the bottom specific information about selected data sets are provided.

Frame re-sizing is available

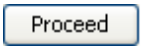
Icons glossary

Common elements are used in ARco with the same behaviour in different contexts:

Load button: Used to upload a data file: gene-expression data in the Transaction tab, a Transactions datafile in the “Frequent itemset tab”; and frequent itemset datafile in the Rules Tab

Browse button: Used to “save” files : transactions, frequent itemsets or rules depending on the tab.

Expand button: Displays the advanced options in the “transactions generation” tab.



Proceed button: Launch the corresponding process.

Next / Previous buttons: browse and surfing the multiple-pages tables.

ARco organization:

Tabs: ARco is organised in four main modules; the natural steps in association rule discovering: procedures:

1. Data manipulation to produce a set of transactions to be mined
2. Finding frequent itemsets in the transactions file
3. Produce association rules
4. Browsing and exploring results

Following these steps the Control Pane has the next tabs or sub-sections:

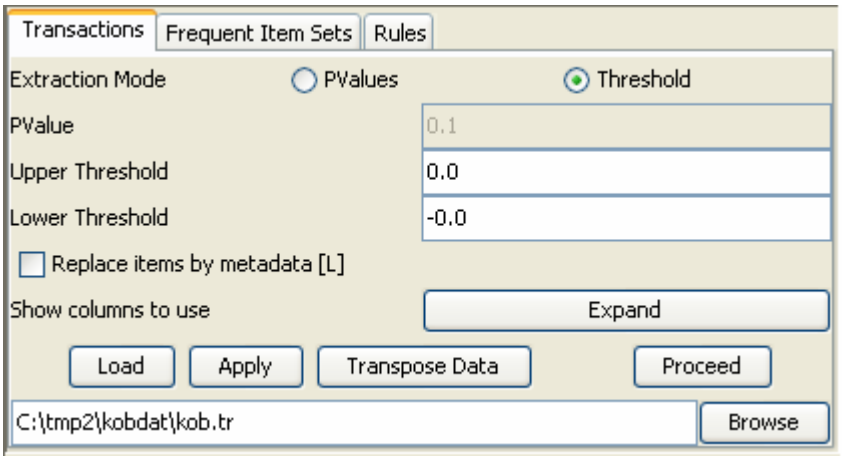
Transactions	filtering , transforming and coding tools to produce transactions
Frequent Item Sets.	Algorithms to produce Frequent Item Sets (k-itemsets: set of k items frequently present together in the same transaction)
Rules	Parameters for association rules production
Data view	Visualization, filtering, translation and exploring rules

Other secondary tabs:

HeatMap	Heatmap representation of expresión data
Histogram	Histogram representation of expresión data
Original Data View	Displays the original Transactions that holds the a selected rule
Visualization Panel	Rule profile display
Info Tabs	Different informative tabs associated with a given action

From original data to Transactions

This section contains the Control Pane with the working options and needed parameters. It contains filtering parameters, items selection and transformation; metadata identifiers, etc.

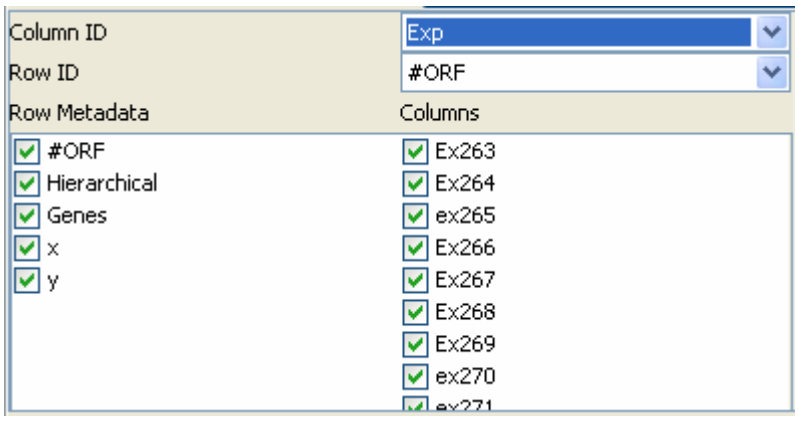


► Parameters

Extraction Mode	Used to transform expression values into 3-state elements: over- and under-expressed and not differentially expressed. Two methods are available: Thresholds (under- and over-) and p-values
pvalue	Maximum p-value to set an expression value as differentially expressed (required when using the *p-value* extraction mode). Under this option, the pvalue associated to each expression ratio will be computed from the z-scores (normalised ratios, with mean zero and standard deviation 1)
Upper Threshold	Over-expression threshold. Minimum expression value to be set as over-expressed (required when using the *threshold* extraction mode)
Lower Threshold	Under-expression threshold. Maximum expression value to be set as under-expressed (required when using the *threshold* extraction mode)
Relpace items by metadata [L]	Instead of including the item-ID in a transaction, it is replaced with the experiment metadata (sample or column metadata).
Apply	Perform the data filtering (using the extraction mode and associated parameters) and up-dates the corresponding images.
Transpose Data	Transpose the matrix (row-columns interchanging). Obviously it includes metadata.

Proceed	Generate transaction from filtered data.

Advanced options are displayed when click on expand button

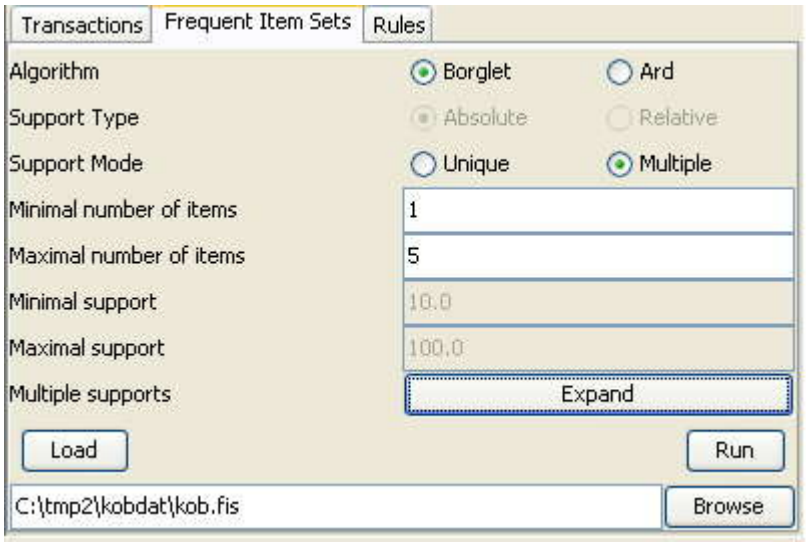


Column ID	Experiment, sample or column identifier
Row ID	Gene or row identifier

In the main body of the dialog box, row and column metadata can be activate/inactivate to participate in the mining procedure.

Frequent Item Set tab

Frequent k-itemset production procedure is controlled from this tab. Main parameter are: Support (number of transactions containing a given k-itemset); maximum k value and algorithm.



Algorithm	Two options are available: Extended (variable support) Borgelt proposal (http://www.borgelt.net/apriori.html) Rodriguez et al. (http://www.biomedcentral.com/1471-2105/7/54)
Support Type	In absolute value (number of transactions) or relative percentage. When working with multiple supports (by item support) this parameter must be specified for each different item.
Support Mode	Unique: the same support for all items Multiple: specific support for each item
Minimal number of items	Minimal k value
Maximal number of items	Maximum k value
Minimal support	Available for Unique support
Maximal support	Available for Unique support

These options and parameters are needed to produce frequent k-itemset with general support. If individual supports are needed for each item we can use the Expand button.

Name	Min support	Max support	Support Type
Hierarchical	10.0	100.0	Relative
Genes	10.0	100.0	Relative
x	10.0	100.0	Relative
y	10.0	100.0	Relative
[+Ex263]	10.0	100.0	Absolute
[+Ex264]	10.0	100.0	Absolute
[+ex265]	10.0	100.0	Absolute
[+Ex266]	10.0	100.0	Absolute

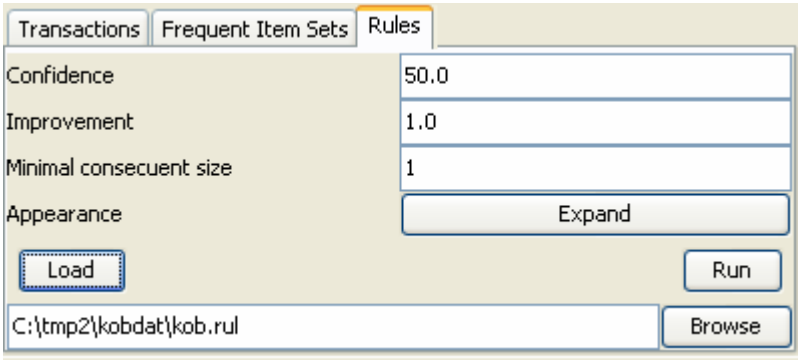
All the item labels are displayed and a dialog box can be used to set individual supports both, modes and values. A table with the following parameters is available:

Name	Item-label (it can correspond to an item, an item metadata or sample metadata).
Min support	Minimum support for this item
Max support	Maximum support for this item
Support Type	Absolute (number of transactions) or relative as percentage.

All values can be modified at the same time using right button functionality. Fir instance, to set to "Absolute" all the Support-type you can click right button over any cell in the "Support-type" column (the same is valid for support values).

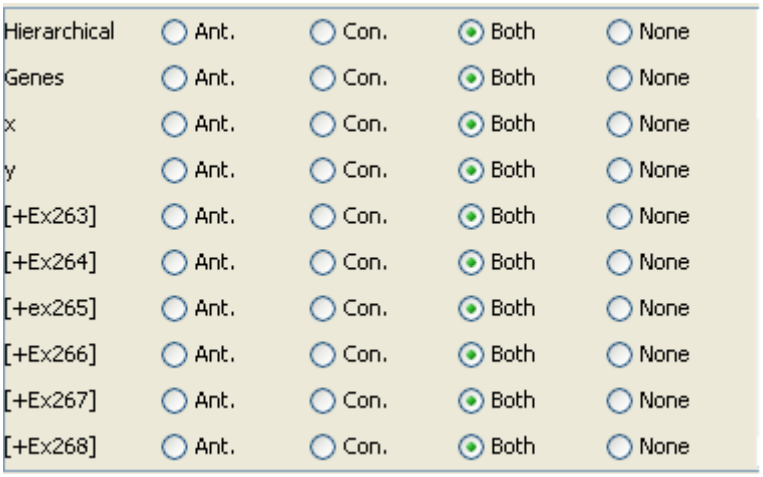
Rules Tab

Option and parameter related with rule production



Confidence	Minimum rule confidence: rule reliability of X => Y in T is the ratio of the # of transactions in T containing X that also contain Y, versus total # of transactions in T containing X to produce a rule
Improvement	Minimum Improvement
Minimal consecuent size	How many items in the consequent side

By default, any element can be at any place in the rule (antecedent or consequent side). Positional restrictions can be established for each item-type to be in the antecedent, in the consequent, in both or not to be in the rule.



Ant.	This data type can only be in the antecedent side of the rule
Con	This data type can only be in the consequent side of the rule
Both	This data type can be both in the antecedent or consequent side of the rule
None	Rules with this datatype are discarded

Data View tab

It becomes available when a data file has been loaded. A table-style is used to display the data set highlighting the cells involved in a transaction production. Some data manipulation tools are available on right-button functionality clicking over the column to be modified.

Data View

#ORF	Hierar...	Genes	x	y	170	153	167	186	188	165	175
				Height	170	153	167	186	188	165	175
				Age	32	22	12	21	11	22	55
				Gender	Male	Female	Female	Female	Male	Male	Female
				Exp	Ex263	Ex264	ex265	Ex266	Ex267	Ex268	Ex269
BG10065	1.2.3 , 2...	dnaA, dn...	1 ,44	1	-0.550197	10.133	-0.838249	0.0687128	168.182	-134.104	-116.993
BG10066	1.2.4 , 2...	dnaN, dn...	2,4	1	-0.499571	181.876	Empty	-0.563901	-0.403897	-207.039	-1.822
BG10077	1.2.3 , 2...	serS	12	1	0.286304	0.910733	-112.553	0.82852	0.906891	-103.953	-117.754
BG10078	1.2.4 , 2...	dck, yaaF	13	1	0.555215	102.975	0	0.347165	-0.765535	-193.587	-119.265
BG10081	1.2.3 , 2...	yaaI	16	1	-168.281	-0.165249	13.516	-118.488	0.247438	-107.039	-0.631355
BG10100	1.2.4 , 2...	abrB, cpsX	36	1	-0.519818	-138.466	0.391579	-0.405712	-0.708446	0.930876	0.10276
BG11565	1.2.3 , 2...	ybbA	28	3	-0.670692	0.0718929	-0.289507	-217.856	-0.396517	-1	-0.10188
BG10837	1.2.4 , 2...	feuC	29	3	-0.0888093	0.367329	-120.256	-154.649	-168.407	-235.908	-0.135655
BG10836	1.2.3 , 2...	feuB	30	3	0.680063	0.64372	-0.698406	-0.969838	0.147104	0.451989	-0.0107264
BG10833	1.2.4 , 2...	ybbC, yzbB	33	3	-141.504	0.0844746	-0.428843	-199.035	-223.704	-112.873	-0.300124
BG10832	1.2.3 , 2...	ybbD, yzbA	34	3	0.294103	-0.563901	-0.0239789	0.159513	-0.744936	-0.150448	-0.00898179
BG11566	1.2.4 , 2...	ybbE	35	3	-0.173016	-0.379237	-0.441705	-0.812067	-192.254	-0.471143	-0.0889005
BG11567	1.2.3 , 2...	ybbF	36	3	0.275634	-0.427862	-0.259387	-0.508791	-0.657475	-0.337695	-0.0313155
BG11569	1.2.4 , 2...	ybbH	37	3	-0.811143	-0.572694	-0.409185	-110.815	-0.805781	-0.35973	-0.51
BG11571	1.2.3 , 2...	ybbJ	39	3	-159.516	-0.613583	-0.415037	-122.408	-119.265	-0.89743	-0.467126
BG11572	1.2.4 , 2...	ybbK	40	3	0.129283	-0.13493	-0.216423	-0.298179	0.165285	-0.171046	-0.440216
BG10166	1.2.3 , 2...	adaA	48	3	0.196397	0.521835	-0.76356	-179.564	-209.311	-182.716	-0.245112
BG10167	1.2.4 , 2...	adaB	49	3	0.257158	104.439	-0.571157	-117.544	0.456378	-0.163499	-0.125531
BG10949	1.2.3 , 2...	ndhF, ybxE	50	3	-0.395138	0.246161	-0.90092	-224.236	-217.864	-204.731	-0.337035

Triming	Keeps the first or last 'n' characters
Hierarchic	Reduce the deep-level value in a hierarchical codification
Interval Equi-Depth	Identify 'n' different groups with equal number of elements (equalization). Valid for numerical data)
Interval Equi-Width	Produce "n" different groups with the same range size. Require min and max values and interval size
Undo Transformation	Un-do the last transformation
Reload Column	Un-do all transformations performed on a given column (re-load original values)

Annexe 1contains detailed information for data transforming procedures

Frequent Items Sets (Visualization) tab

Displays the frequent item sets. It can be explored and ordered by the item support (absolute or relative)

Frequent Item Sets

Set	Support	Number of transactions
[-ex274]	1,80	13,00
[-Ex264]	2,00	14,00
[+ex273]	2,40	17,00
[+ex285]	2,00	14,00
[+ex270]	2,10	15,00
[-ex279]	3,70	26,00
[+ex272]	2,70	19,00
[-Ex267]	2,10	15,00
[+ex286]	2,40	17,00
[+ex280]	4,40	31,00
[+ex283]	3,10	22,00
[+ex276]	3,40	24,00
[+ex284]	4,70	33,00
[+ex279]	6,80	48,00
[-Ex268]	4,90	35,00
[-ex270]	5,20	37,00
[-Ex266]	11,00	78,00
[+ex283] [+ex286]	2,10	15,00
[-Ex266] [-Ex268]	2,80	20,00
[-Ex268] [-ex271]	3,20	23,00
[-Ex266] [-ex270]	2,30	16,00
[-ex270] [-ex271]	4,10	29,00
[-Ex266] [-ex271]	8,10	57,00
[-Ex266] [-Ex268] [-ex271]	2,10	15,00

Rules (Visualization) tab.

Tab used to display rules. Rules can be ordered by any of their numeric columns.

Rules

Antecedent	Consecuent	Confidence	Support	ABS Support	Coverage	Improvem...	Leverage	Conviction	Entropy	RuleID
[+181]	[-155]	78,16	9,60	68,00	12,29	1,79	4,22	257,40	0	0
[+186]	[-186]	56,10	9,75	69,00	17,37	1,22	1,77	123,22	0	1
[-170]	[-186]	88,24	12,71	90,00	14,41	1,92	6,10	459,82	0	2
[+181]	[-186]	73,56	9,04	64,00	12,29	1,60	3,40	204,62	0	3
[-176]	[-155]	52,94	5,08	36,00	9,60	1,21	0,88	119,46	0	4
[-176]	[-188]	62,90	5,51	39,00	8,76	2,93	3,63	211,69	0	5
[-196] [-176]	[-155]	90,00	5,08	36,00	5,65	2,06	2,61	562,15	0	6
[-155] [-176]	[-196]	80,00	5,08	36,00	6,36	3,06	3,42	369,35	0	7
[-176]	[-186]	72,58	6,36	45,00	8,76	1,58	2,34	197,29	0	8
[-196]	[-155]	89,73	23,45	166,00	26,13	2,05	12,01	547,35	0	9
[-155]	[-196]	53,55	23,45	166,00	43,79	2,05	12,01	159,03	0	10
[-186] [-176]	[-155]	91,11	5,79	41,00	6,36	2,08	3,01	632,42	0	11
[-155] [-176]	[-186]	91,11	5,79	41,00	6,36	1,98	2,87	608,58	0	12
[-176]	[-186]	57,35	5,51	39,00	9,60	1,25	1,10	126,85	0	13
[-165]	[-155]	91,48	22,74	161,00	24,86	2,09	11,86	659,59	0	14
[-155]	[-165]	51,94	22,74	161,00	43,79	2,09	11,86	156,33	0	15
[-188]	[-155]	80,92	17,37	123,00	21,47	1,85	7,97	294,64	0	16
[-165]	[-196]	64,77	16,10	114,00	24,86	2,48	9,61	209,70	0	17
[-196]	[-165]	61,62	16,10	114,00	26,13	2,48	9,61	195,79	0	18
[-188]	[-196]	66,45	14,27	101,00	21,47	2,54	8,66	220,16	0	19
[-196]	[-188]	54,59	14,27	101,00	26,13	2,54	8,66	172,96	0	20

E=>E E=>M M=>E M=>M Show all Filters << [0 - 134] >>

Clicking a given rule, all transactions that hold the rule are highlighted in the “Data frame”.

Different filters are available

E=>E	Experiment – Experiment rules (only expresión values)
E=>M	Experiment values (antecedent) implies a Metadata (consequent)
M=>E	Metadata in the antecedent and experiment value in the consequent
M=>M	Metadata – Metadata rules
Show all	Show all rules
Filters	Advanced filters

Hide trivials

Remove trivial rules. A rule is **trivial** if there is another rule with the same Right-Hand-Side and a subset of the Left-Hand-Side that covers exactly the same cases from the data set. For example, the first of the two rules below is trivial because it has the same coverage as the second. Adding Tomatoes to the LHS of the second rule does not affect it.

Lettuce & Tomatoes -> Cucumber [Coverage=0.250 (250); Support=0.239 (239); Strength=0.956; Lift=2.91; Leverage=0.1568 (156)]

Lettuce -> Cucumber [Coverage=0.250 (250); Support=0.239 (239); Strength=0.956; Lift=2.91; Leverage=0.1568 (156)]

If a rule is trivial then it will have the same support, strength, lift, and leverage as the rule with respect to which it is trivial.

(see http://www.rulequest.com/MOfiltering.html)

Hide unproductive

Unproductive rules. A rule is **unproductive** if there is another rule with the same Right-Hand-Side and a subset of the Left-Hand-Side that has equal or higher strength. For example, the first of the rules below is unproductive because it has lower strength than the second. Adding Promotion1=f to the LHS of the second rule decreases its performance.

Profitability99 < 419 & Promotion1=f -> Spend99 < 2030 [Coverage=0.274 (274); Support=0.248 (248); Strength=0.905; Lift=2.72; Leverage=0.1568 (156)]

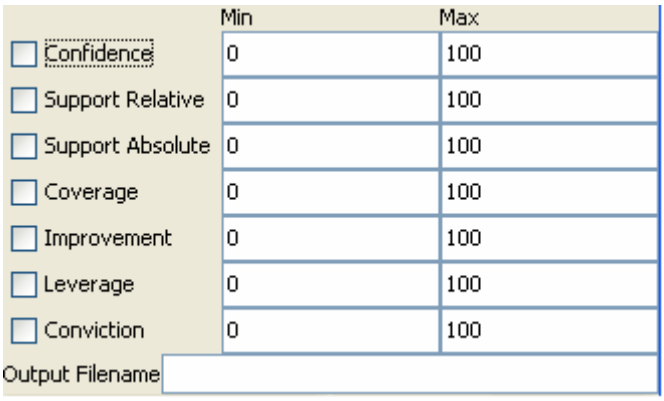
Profitability99 < 419 -> Spend99 < 2030 [Coverage=0.333 (333); Support=0.302 (302); Strength=0.907; Lift=2.72; Leverage=0.1911 (191)]

If a rule is unproductive then it will have the same or worse support, strength, lift, and leverage as the rule with respect to which it is unproductive.

(see http://www.rulequest.com/MOfiltering.html)

Custom Filter — Customised filter

Customised filters allow combining different requirements to filter the rules. In the available dialog-box several criterions can be used at the same time (confidence, support, coverage, etc.) and minimum and maximum values must be specified for each criterion. Output is stored in a file

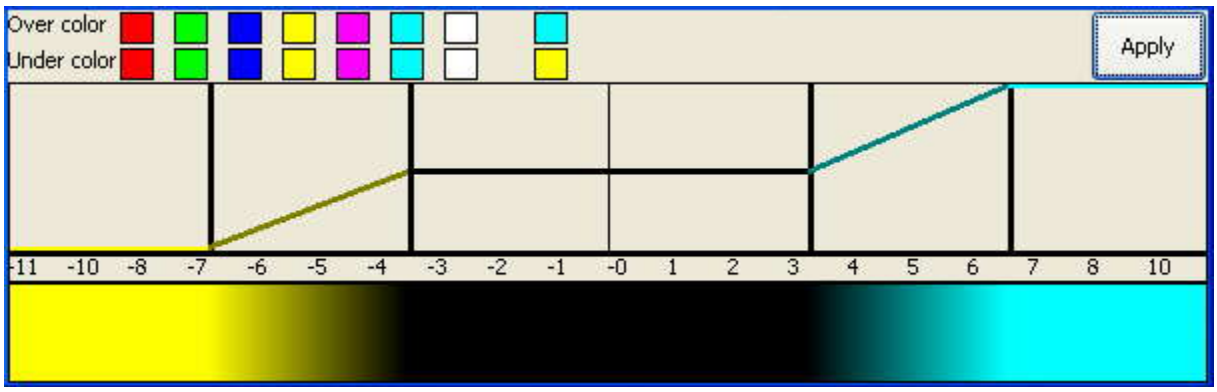


Heat-Map tab

This frame is used to display a visual representation of gene expression values in the form of a coloured matrix. Traditionally expression values have been represented using red for over-expression and green for under-expressed genes. The colour scale also includes a “black” range for values (log2 ratios) close to zero, and red and green scale for different values, including a saturation point from which all the values receive the same colour.

Image can be saved to disk using right-button: “Save image” functionality.

The colour palette and saturation points are customisable. Right-button: Colours Palette

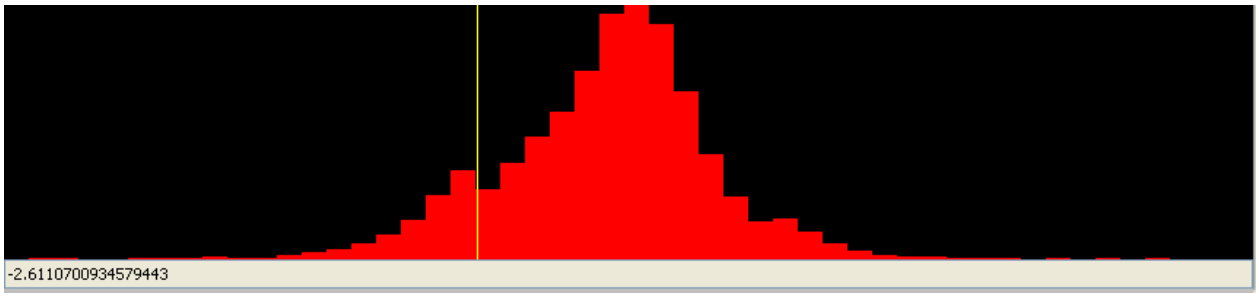


Over and under- colours are used for over-expressed and under-expressed genes. Changes affect Data view representation

In the main body there are 4 vertical lines that can be horizontally moved to define the “non differentially expressed range –around log ratio equal cero)” and under / over expressed points at which the signal become saturated (al values at the left in the under-expression side or all values on the right of the saturation points are coded with the same colour).

Histogram tab

Histogram of gene-expression values (original data).



Numerical values are shown on the bottom bar when the mouse moves over the image.

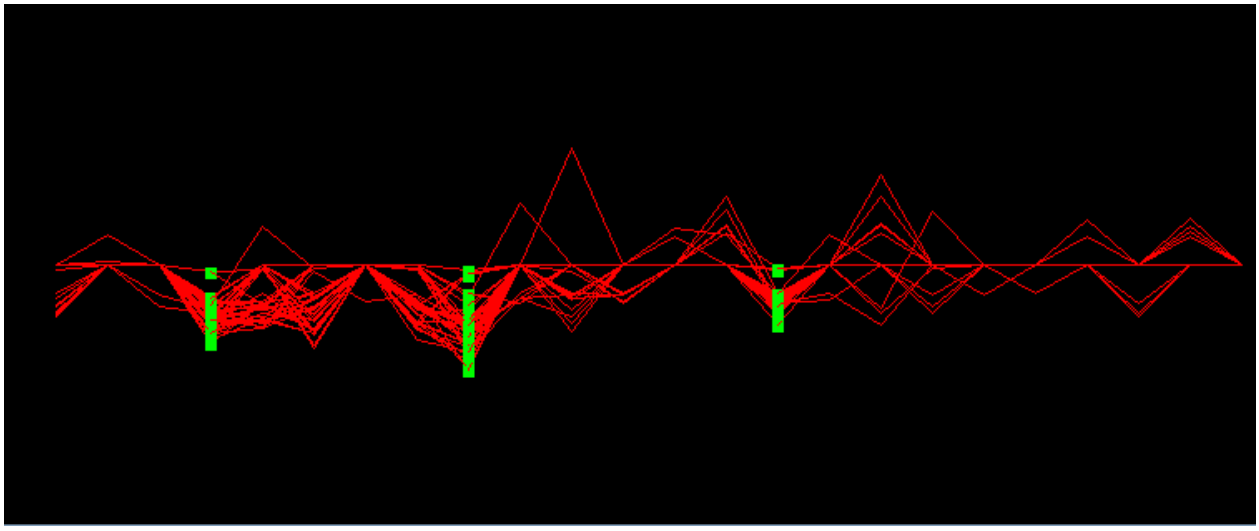
Original Data View tab

This frame is used to display the transactions that hold the rule (selected rule) and only will be availbale after the rule selection event.

#ORF	Hierarchical	Genes	x	y	170	153	167	186	188	165	
				Height	170	153	167	186	188	165	1
				Age	32	22	12	21	11	22	5
				Gender	Male	Female	Female	Female	Male	Male	F
				Exp	Ex263	Ex264	ex265	Ex266	Ex267	Ex268	E
BG10081	1.2.3 , 2.3.5	yaaI	16	1	-168.281	-0.165249	13.516	-118.488	0.247438	-107.039	-
BG10833	1.2.4 , 2.3.7	ybbC, yzbB	33	3	-141.504	0.0844746	-0.428843	-199.035	-223.704	-112.873	-
BG11571	1.2.3 , 2.3.8	ybbJ	39	3	-159.516	-0.613583	-0.415037	-122.408	-119.265	-0.89743	-
BG12711	1.2.3 , 2.3.10	ybcL	56	3	-170.044	0.274507	-0.303781	-213.993	-196.963	-247.249	-
BG11504	1.2.4 , 2.3.12	csgA	12	4	-119.265	0.539159	0.0	-198.793	-147.089	-0.793549	-
BG10173	1.2.4 , 2.3.19	ycxB	30	6	-147.732	0.0792633	-0.053638	-188.604	-124.691	-0.527247	-
BG12053	1.2.3 , 2.3.21	ydaE	31	7	-196.963	0.184233	138.215	-12.555	-0.406424	-135.669	-
BG12067	1.2.3 , 2.3.22	ydaT	49	7	-119.465	0.12063	116.046	-148.843	0.201634	-144.057	-
BG12804	1.2.3 , 2.3.30	ydjN	46	10	-130.812	0.461179	0.571313	-167.807	-0.464963	-133.985	-
BG12841	1.2.3 . 2.3.33	verO	27	11	-129.399	-0.31034	-0.304855	-139.734	10.889	0.0687128	-

Visualization tab

For E-E rules, displays the gene-expression profile (or the sample-profile for transposed matrices) with red lines; and the experiments that hold the rule (green boxes)



Options to modify the representation are available on right-buttom: “Change View”

Show backgroundgrid	Displays a grid
Show all profiles	OFF: displays the gene-expression-profile of those genes holding the rule. ON: displays all the gene-expression profiles as a background image, the gene-expression profile of those genes holding the rule coloured in the foreground and the green boxes for items involved in the rule
Draw lines/Draw dots	Displays only the green-boxes or also draws a “rule profile” (joint with a line all the points)

Filtering rules

Antecedent	Consecuent	Confidence	Support	ABS Support	Coverage	Improvement	Leverage	Conviction	Entropy	RuleID
[Economical le...	[PostalCode.2...	66,67	16,00	16,00	24,00	2,08	8,32	204,00	0	0
[PostalCode.2...	[Economical le...	50,00	16,00	16,00	32,00	2,08	8,32	152,00	0	1
[Economical le...	[PostalCode.p...	61,36	27,00	27,00	44,00	1,43	8,08	147,53	0	2
[PostalCode.p...	[Economical le...	62,79	27,00	27,00	43,00	1,43	8,08	150,50	0	3
[Economical le...	[PostalCode.p...	61,90	13,00	13,00	21,00	2,48	7,75	196,88	0	4
[PostalCode.p...	[Economical le...	52,00	13,00	13,00	25,00	2,48	7,75	164,58	0	5
[PostalCode.2...	[Civilstage.ma...	53,13	17,00	17,00	32,00	1,15	2,28	115,20	0	6
[PostalCode.p...	[Civilstage.sin...	55,81	24,00	24,00	43,00	1,12	2,50	113,16	0	7
[Economical le...	[Civilstage.ma...	70,83	17,00	17,00	24,00	1,54	5,96	185,14	0	8
[Economical le...	[PostalCode.2...	76,47	13,00	13,00	17,00	2,39	7,56	289,00	0	9
[PostalCode.2...	[Economical le...	76,47	13,00	13,00	17,00	3,19	8,92	323,00	0	10
[PostalCode.2...	[Civilstage.ma...	81,25	13,00	13,00	16,00	1,77	5,64	288,00	0	11
[Civilstage.sin...	[Economical le...	50,00	25,00	25,00	50,00	1,14	3,00	112,00	0	12
[Economical le...	[Civilstage.sin...	56,82	25,00	25,00	44,00	1,14	3,00	115,79	0	13
[PostalCode.2...	[Economical le...	66,67	10,00	10,00	15,00	1,52	3,40	168,00	0	14
[PostalCode.2...	[Civilstage.sin...	71,43	10,00	10,00	14,00	1,43	3,00	175,00	0	15
[+champagne]	[PostalCode.2...	57,14	32,00	32,00	56,00	1,79	14,08	158,67	0	16
[PostalCode.2...	[+champagne]	100,00	32,00	32,00	32,00	1,79	14,08	Infinity	0	17
[+beer]	[PostalCode.p...	64,00	16,00	16,00	25,00	2,56	9,75	208,33	0	18
[PostalCode.p...	[+beer]	64,00	16,00	16,00	25,00	2,56	9,75	208,33	0	19
[Economical le...	[PostalCode.p...	60,00	15,00	15,00	25,00	1,40	4,25	142,50	0	20
[PostalCode.p...	[Economical le...	62,50	15,00	15,00	24,00	1,42	4,44	149,33	0	21
[PostalCode.p...	[Civilstage.sin...	55,56	15,00	15,00	27,00	1,11	1,50	112,50	0	22
[PostalCode.2...	[+wine High Q]	100,00	32,00	32,00	32,00	1,33	8,00	Infinity	0	23
[Economical le...	[Civilstage.ma...	52,38	11,00	11,00	21,00	1,14	1,34	113,40	0	24

E=>E E=>M M=>E M=>M Show all Filters << [0 - 306] >>

Filtering by datatype: E = Experiments; M = Metadata

Experiment Rules: Antecedent and Consequent are expresión values

Metadata Rules: Antecedent and Consequent are metadata values

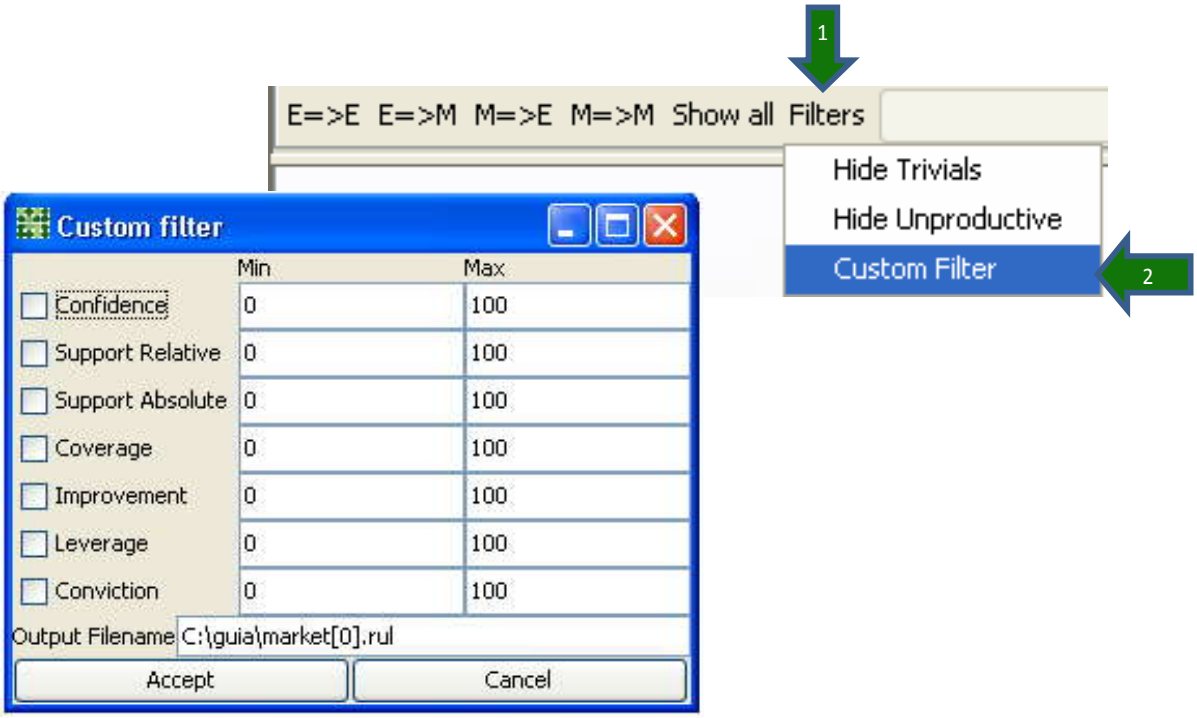
E=>E E=>M M=>E M=>M Show all Filters

Experiment⇒Metadata: the antecedent is an expression value and the consequent is a metadata

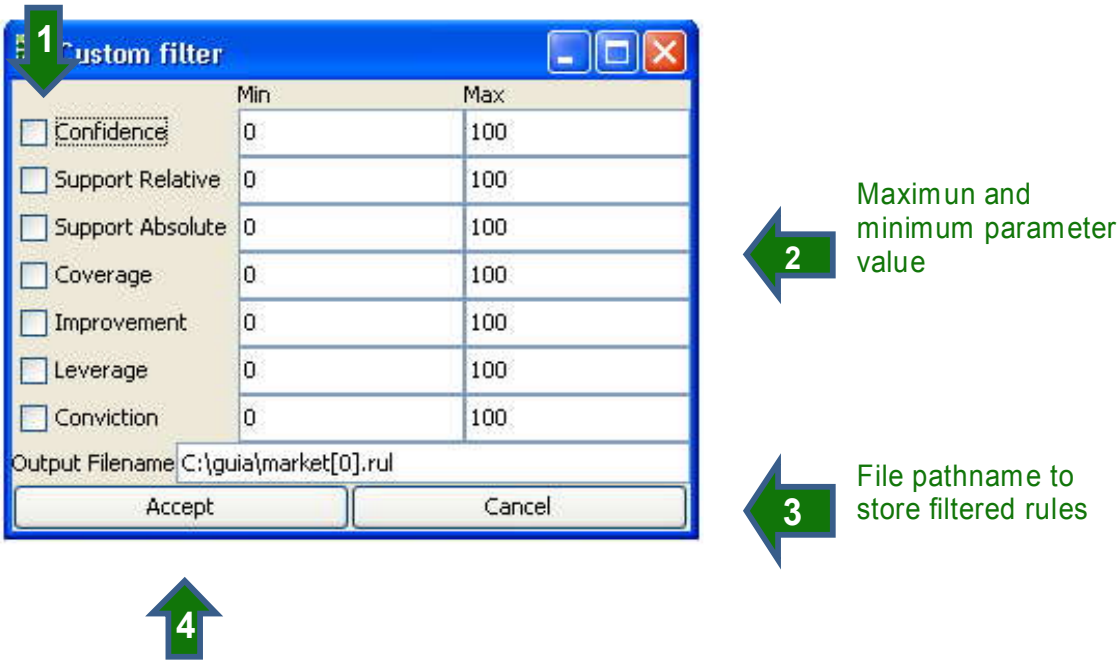
Metadata⇒Experiment: the antecedent is a metadata and the consequent is an expression value

Shows all the rules

Filtering by values



Set filtering parameters



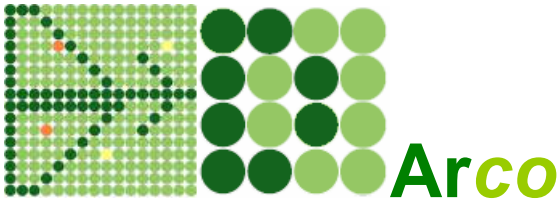
Visualization of transactions that hold the rule

Antecedent	Consecuent	Confidence	Support	ABS Sup...	Coverage	Improve...	Lever...	Conviction	Entropy	RuleID
[Economical level(ABCD).D]	[PostalCode.29001]	66,67	16,00	16,00	24,00	2,08	8,32	204,00	0	0
[PostalCode.29001]	[Economical level(...	50,00	16,00	16,00	32,00	2,08	8,32	152,00	0	1
[Economical level(ABCD).C]	[PostalCode.pc29...	61,36	27,00	27,00	44,00	1,43	8,08	147,53	0	2
[PostalCode.pc29002]	[Economical level(...	62,79	27,00	27,00	43,00	1,43	8,08	150,50	0	3
[Economical level(ABCD).B]	[Posta .pc29...	61,90	13,00	13,00	21,00	2,48	7,75	196,88	0	4
[PostalCode.pc29003]	[Econ level(...	52,00	13,00	13,00	25,00	2,48	7,75	164,58	0	5
[PostalCode.29001]	[Civil rried]	53,13	17,00	17,00	32,00	1,15	2,28	115,20	0	6
[PostalCode.pc29002]	[Civilstage.single]	55,81	24,00	24,00	43,00	1,12	2,50	113,16	0	7
[Economical level(ABCD).D]	[Civilstage.married]	70,83	17,00	17,00	24,00	1,54	5,96	185,14	0	8
[Economical level(ABCD).D] ...	[PostalCode.29001]	76,47	13,00	13,00	17,00	2,39	7,56	289,00	0	9
[PostalCode.29001] [Civilst...	[Economical level(...	76,47	13,00	13,00	17,00	3,19	8,92	323,00	0	10
[PostalCode.29001] [Econo...	[Civilstage.married]	81,25	13,00	13,00	16,00	1,77	5,64	288,00	0	11
[Civilstage.single]	[Economical level(...	50,00	25,00	25,00	50,00	1,14	3,00	112,00	0	12
[Economical level(ABCD).C]	[Civilstage.single]	56,82	25,00	25,00	44,00	1,14	3,00	115,79	0	13
[PostalCode.29001] [Civilst...	[Economical level(...	66,67	10,00	10,00	15,00	1,52	3,40	168,00	0	14
[PostalCode.29001] [Econo...	[Civilstage.single]	71,43	10,00	10,00	14,00	1,43	3,00	175,00	0	15

E=>E E=>M M=>E M=>M Show all Filters << [0 - 306] >>

Original Data View

CustomerID	PostalCode	Economic...	Civilstage	A1	A2	A1	A1	A2
			Supplier	A1	A2	A1	A1	A2
			Price range	P1	P1	P2	P2	P1
			Restrictions	Yes	Not	Yes	Yes	Yes
			Exp	beer	juice	champagne	wine High Q	wine X
49	pc29002	D	married	Empty	Empty	3.0	3.0	Empty
50	pc29002	D	married	Empty	Empty	3.0	3.0	Empty
65	pc29002	D	married	Empty	3.0	Empty	3.0	Empty
66	pc29002	D	married	Empty	4.0	4.0	4.0	Empty
69	29001	D	married	Empty	4.0	4.0	4.0	Empty
70	29001	D	married	Empty	4.0	4.0	4.0	Empty
73	29001	D	married	Empty	4.0	4.0	4.0	Empty
74	29001	D	married	Empty	4.0	4.0	4.0	Empty
92	29001	D	married	Empty	Empty	4.0	4.0	Empty
93	29001	D	married	Empty	Empty	4.0	4.0	Empty
94	29001	D	married	Empty	Empty	4.0	4.0	Empty
95	29001	D	married	Empty	Empty	4.0	4.0	Empty



a Bitlab software

Association Rules collaborative tool

Integrated suite for association rule discovering in medical and molecular data

---

Annexes

Version v1: 8th November 2007.
On-line updated information available at:
http://chirimoyo.ac.uma.es/arco

Comments to: ots@ac.uma.es

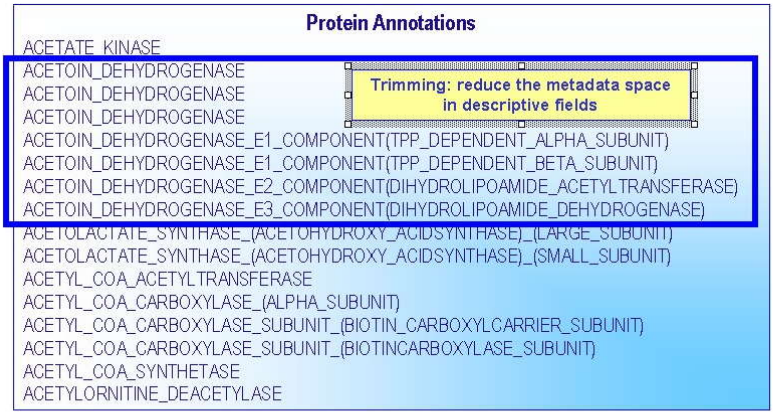
Annexe 1: Data transformation tools

Before producing transactions, it is possible to perform some data transformations with the aim to increase the probability of discovering new knowledge. For instance, if we have extremely descriptive metadata (e.g. patient age) it will be difficult to incorporate this metadata in a frequent itemset. Therefore it could be better to define some categories or groups with similar metadata (e.g. age ranges: {0-10; 11-20; 21-40; 41-60; 61-85; 86-+}).

After rule generation to allow a better analysis, the original data are shown in the “Original Data View” tab.

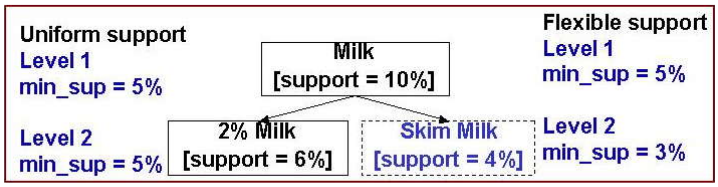
The following transformations are available for metadata:

Trimming: Kept the first T characters.



Example: Trimming the 20 first characters will code the boxes metadatas into the general “ACETOIN_ DEHYDROGENAS”.

Hierarchical:
Items are often organised in hierarchical way, and some transformations also produce hierarchical data. This characteristic has effect in the expected support of item (items at the lower level are expected to have lower support). Since, some fields of the transactions database have this structure; ARco provides a way to re-code the level at which the metadata are annotated



Distance-based data transformations (Interval Equi-Depth & equi-width)

Price	Equi-width	Equi-depth	Distance-based
7	[0,10]	[7,20]	[7,7]
20	[11,20]	[22,50]	[20,22]
22	[21,30]	[51,53]	
50	[31,40]		
51	[41,50]		[50,53]
53	[51,60]		

Incomplete data : Filling

Binning methods do not always capture semantic of data intervals. In these situations a distance-based partitioning can be used, this is to say, numeric attributes can be dynamically discretised to maximise the confidence or compactness of the rules. This discretization can be done by accounting the number of points in a given interval or by “closeness” of points in an interval (distance function)

Triming

Keep the first or last 'n' characters and it is available for all data types (all datatypes are taken as string data).

ID
ACETATE_KINASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE_E1_COMPONENT(TPP_DEPENDENT_ALPHA_SUBUNIT)
ACETOIN_DEHYDROGENASE_E1_COMPONENT(TPP_DEPENDENT_BETA_SUBUNIT)
ACETOIN_DEHYDROGENASE_E2_COMPONENT(DIHYDROLIPOAMIDE_ACETYLTRANSFERASE)
ACETOIN_DEHYDROGENASE_E3_COMPONENT(DIHYDROLIPOAMIDE_DEHYDROGENASE)
ACETOLACTATE_SYNTHASE_(ACETOHYDROXY_ACIDSYNTHASE)_(LARGE_SUBUNIT)
ACETOLACTATE_SYNTHASE_(ACETOHYDROXY_ACIDSYNTHASE)_(SMALL_SUBUNIT)

In this example we can devise three main groups of annotations with at different level of detail. One of the main purposes of data transformation is to increase the probability of a given itemset to be part of a frequent itemset. However, a disperse space of metadata can go on the converse direction. This transformation allows to joint similar data under the same category increasing the support of the categories

ID
ACETATE_KINASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOIN_DEHYDROGENASE
ACETOLACTATE_SYNTHASE
ACETOLACTATE_SYNTHASE

In the example, the first 21 characters on the left are used to describe the category. As result, the descriptor keeps the main power, but additionally, several items will contain it.

Important note: When a trimmed item is part of a rule, ARco will display the original value of the item.

Hierarchical

This type of data transformation is used to reduce the deep level in a hierarchical metadata (the metadata must be in the form of {XsYsZ, where X, Y and Z are a category and 's' is a separator

genName	Functional category	FC_level3
aceE	Metabolism	4
aceF	pyruvate dehydrogenase E2 component	4.2.1
ackA	acetate kinase	4.2
acpP	carrier protein	4.5.1
acpS	holo-[acyl-carrier protein] synthase	4.5.1
adk	adenylate kinase	4.6.2
ahpC	hydroperoxide reductase, C22 subunit, thioredoxin-like	3.3
alaS	alanyl-tRNA synthetase	1.2.2
amiB	Peptidoglycan biosynthesis	5.2
ansA	L-asparaginase I	4.4.2
apaH	diadenosine tetraphosphatase	4.6.2
apbE	thiamine biosynthesis lipoprotein ApbE precursor	4.8.11

In the example, the functional category of genes is shown togeter with the geneName in the first row and the numeroc level of the category. The deper the level is, the more specific the description is. Reduce specificity can be obtained by Hierarchical transformation

In this case we set a retdution to the second level (those annotation whose original category is lower than 2 maintain their initial values).

Result of data transformation are displayed in the picture

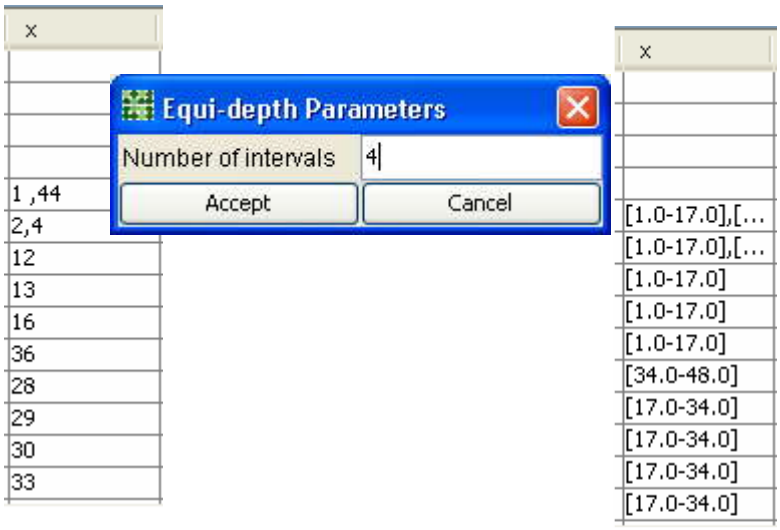
genName	Functional category	FC_level3
aceE	Metabolism	4
aceF	pyruvate dehydrogenase E2 component	4.2
ackA	acetate kinase	4.2
acpP	carrier protein	4.5
acpS	holo-[acyl-carrier protein] synthase	4.5
adk	adenylate kinase	4.6
ahpC	hydroperoxide reductase, C22 subunit, thioredoxin-like	3.3
alaS	alanyl-tRNA synthetase	1.2
amiB	Peptidoglycan biosynthesis	5.2
ansA	L-asparaginase I	4.4
apaH	diadenosine tetraphosphatase	4.6
apbE	thiamine biosynthesis lipoprotein ApbE precursor	4.8

Data categorization

These transformations allow to group numerical values into a reduced set of categories (partitioning). The next options are available:

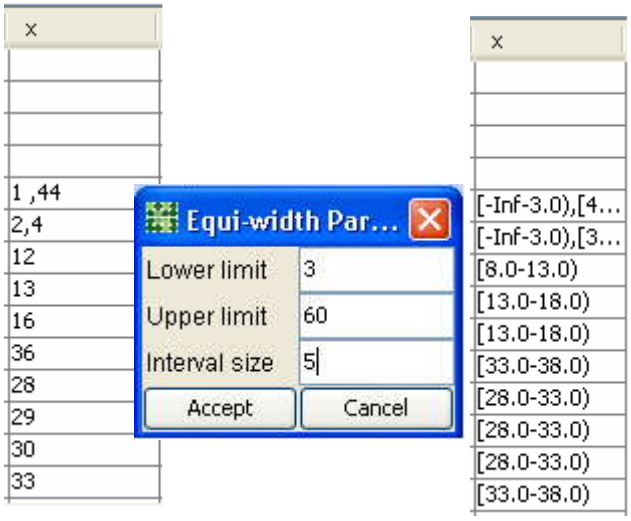
Equi-Depth interval transformation.

Each partition (interval) has the same number of items.



In the example 4 groups have been created with similar number of elements

Equi-Width intervals transformation



Each partition (interval) has the same size (range and interval sizes are required)

Groups values between 3 and 60 in intervals of size 5

Annexe 2: File formats

engene© format (*.dat) http://chirimoyo.ac.uma.es/engenet)

An **engene** data file is a table. This table is stored in the file as a set of fields separated by TAB, and along several lines. This text format may be worked out by Excel. So, an Excel table as follows will generate a file as shown below, when it is saved-as-text.

	A	B	C
1	1	16	72
2	123	15	1
3	151	15	23
4	32	516	53

```
1       16      72
123     15      1
151     15      23
32      516     53
```

	A	B	C
1	NAN	16	72
2	123	15	UNK
3	151		23
4	32	516	53

Data are a collection of vectors, one vector a row. All vectors have the same number of *variables*, one variable a column. Some values may be unknown; in this case, the respective field may be a non numeric string o may be null. These values are called NaN (Not A Number). In the picture, these values are red marked.

It is possible to append notes to data. This kind of information is called metadata. There are three types of metadata: global labels, row labels and column labels. All labels have two parts: the label_name and the label_values. For each global labels name there is only one-value.Row labels have one-value for each data-row; and column labels have one_value for each data column. Next picture shows how to put labels to data.

	A	B	C	D	E
1		CTag1Name	CTag1Val1	CTag1Val2	CTag1Val3
2		CTag2Name	CTag2Val1	CTag2Val2	CTag2Val3
3	RTag1Name				
4	RTag1Val1		1	16	72
5	RTag1Val2		123	15	1
6	RTag1Val3		151	15	23

e red, and the
abels names
e blue. Global
d the values

etween
pace in next
ave fields
v and column
n next figure),
ing after the
in the figure).

	A	B	C	D	E
1		CTag1Name	CTag1Val1	CTag1Val2	CTag1Va
2		CTag2Name	CTag2Val1	CTag2Val2	CTag2Va
3	RTag1Name				
4	RTag1Val1		1	16	
5	RTag1Val2		123	15	
6	RTag1Val3		151	15	

Excel, you must
on used to
ne works with
ds separation and
decimal separator

	A	B	C	D	E	F	G	H	I	J
1					Supplier	A1	A2	A1	A1	A2
2					Price range	P1	P1	P2	P2	P1
3					Restrictions	Yes	Not	Yes	Yes	Yes
4					Exp	beer	juice	champagne	wine High Q	wine X
5	CustomerID	PostalCode	Economical I	Civilstage						
6	1	pc29003	A	Single		2	2			2
7	2	pc29003	A	Single		2	2			2
8	3	pc29003	a	single		2	2			2
9	4	pc29003	a	single		2	2			2
10	5	pc29003	B	single		2	2			2
11	6	pc29003	B	single		2	2	2	2	
12	7	pc29003	B	single		2	3			
13	8	pc29003	B	single		2	2		2	
14	9	pc29003	B	single		2	2		2	
15	10	pc29003	A	single		2	2		2	
16	11	pc29003	A	single		2	2		2	
17	12	pc29003	a	single		2	2		2	
18	13	pc29003	A	single		2	3		2	
19	14	pc29003	A	Married			2		2	
20	15	pc29003	C	Married			2		2	
21	16	pc29003	C	married			2			2
22	17	pc29003	C	married			2			2
23	18	pc29003	B	married			2	2	2	
24	19	pc29003	B	married		2	2	-2		
25	20	pc29003	B	married			4		2	
26	21	pc29003	B	married			4			2
27	22	pc29003	B	married		2	4			
28	23	pc29003	B	married			2	2	2	
29	24	pc29003	B	married		2	4			
30	25	pc29003	B	married			2	2	2	

Gene (or row) Metadata labels are shown in red and data in orange. Dark green represents sample (or experiment / column) metadata labels; and light green column metadata values. Gene-expression ratios are shown in blue.

The same format is valid as *.xls (Excel file); or TSV –text tabulated- / CVS –comma separated) files

	A	B	C	D	E	F	G	H	I
1				Supplier	A1	A2	A1	A1	A2
2				Price range	P1	P1	P2	P2	P1
3				Restrictions	Yes	Not	Yes	Yes	Yes
4	CustomerID	PostalCode	Economical I	Civilstage	beer	juice	champagne	wine High Q	wine X
5	1	pc29003	A	Single	2	2			2
6	2	pc29003	A	Single	2	2			2
7	3	pc29003	a	single	2	2			2
8	4	pc29003	a	single	2	2			2
9	5	pc29003	B	single	2	2			2
10	6	pc29003	B	single	2	2	2	2	
11	7	pc29003	B	single	2	3			
12	8	pc29003	B	single	2	2		2	
13	9	pc29003	B	single	2	2		2	
14	10	pc29003	A	single	2	2		2	
15	11	pc29003	A	single	2	2		2	
16	12	pc29003	a	single	2	2		2	
17	13	pc29003	A	single	2	3		2	
18	14	pc29003	A	Married		2		2	
19	15	pc29003	C	Married		2		2	
20	16	pc29003	C	married		2			2
21	17	pc29003	C	married		2			2
22	18	pc29003	B	married		2	2	2	
23	19	pc29003	B	married	2	2	-2		
24	20	pc29003	B	married		4		2	
25	21	pc29003	B	married		4			2
26	22	pc29003	B	married	2	4			
27	23	pc29003	B	married		2	2	2	
28	24	pc29003	B	married	2	4			
29	25	pc29003	B	married		2	2	2	
30	26	pc29002	B	single	3				3
31	27	pc29002	A	single	3				3
32	28	pc29002	C	single	3				3
33	29	pc29002	C	single	3				3
34	30	pc29002	C	single		4	3	3	
35	31	pc29002	B	single		4	3	3	
36	32	pc29002	B	married		4	3	3	

Gene (or row) Metadata labels are shown in red and data in orange. Dark green represents sample (or experiment / column) metadata labels; and light green column metadata values. Gene-expression ratios are shown in blue.