Association Rule Mining on RNAseq Data

Trends in Bioinformatics

Lukas Ehrig

Jan 23th, 2019

Agenda

1. **Motivation**
2. **Method**
   - 2.1. Filtering Association Rules
   - 2.2. Interestingness Measures
   - 2.3. Weighting of Interestingness Measures
3. **Experiments**
4. **Conclusion**

1. Motivation

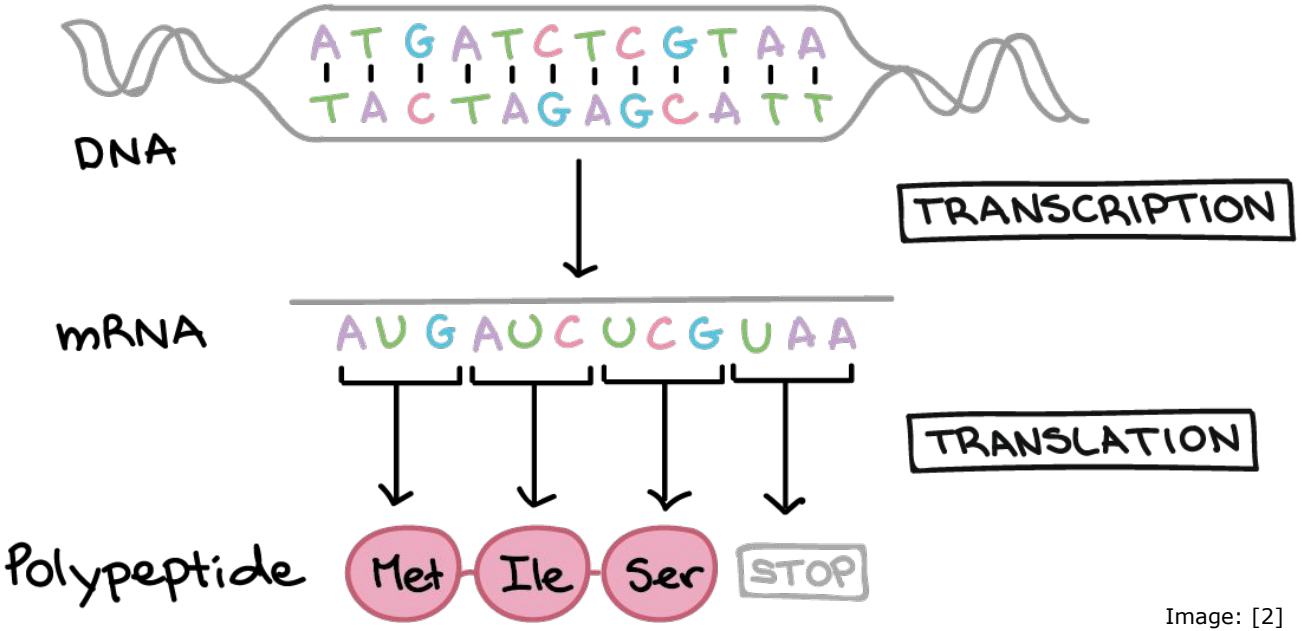


Image: [2]

1. Motivation

Association Rule Mining

Discovering significant relations between attributes (“items”) in large datasets.

Association Rules

Gene1↑ → Gene2↓

Illness → Gene1↑, Gene2↓

Gene1↑, Gene2↓ → Illness

	Gene1	Gene2	Gene3	Gene4	…
Sample 1	↑	↓		↑	…
Sample 2		↓			…
Sample 3		↑	↑		…
…	…	…	…	…	…

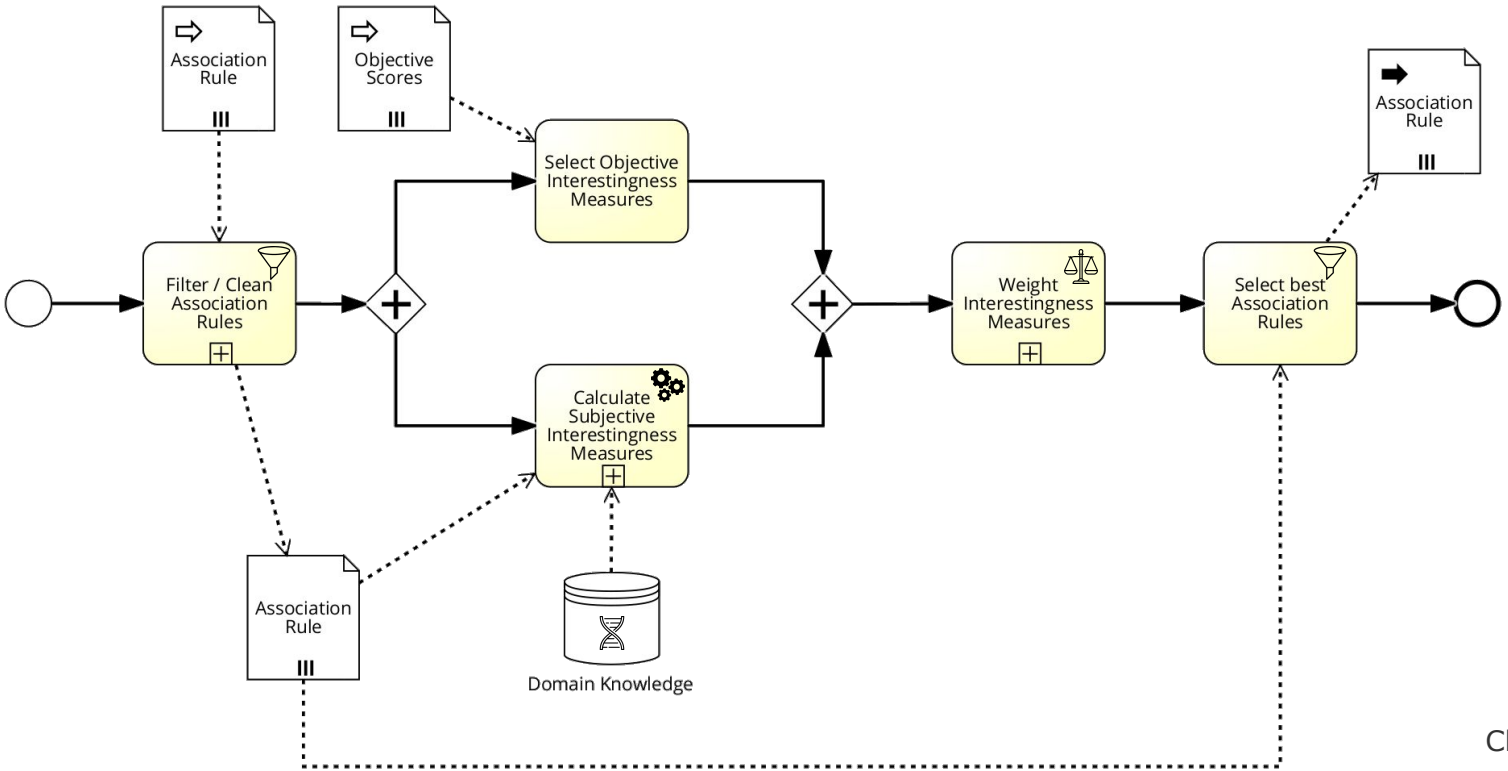
1. Motivation

Number of features infeasible
Exponential number of possible frequent item sets
→ Runtime complexity

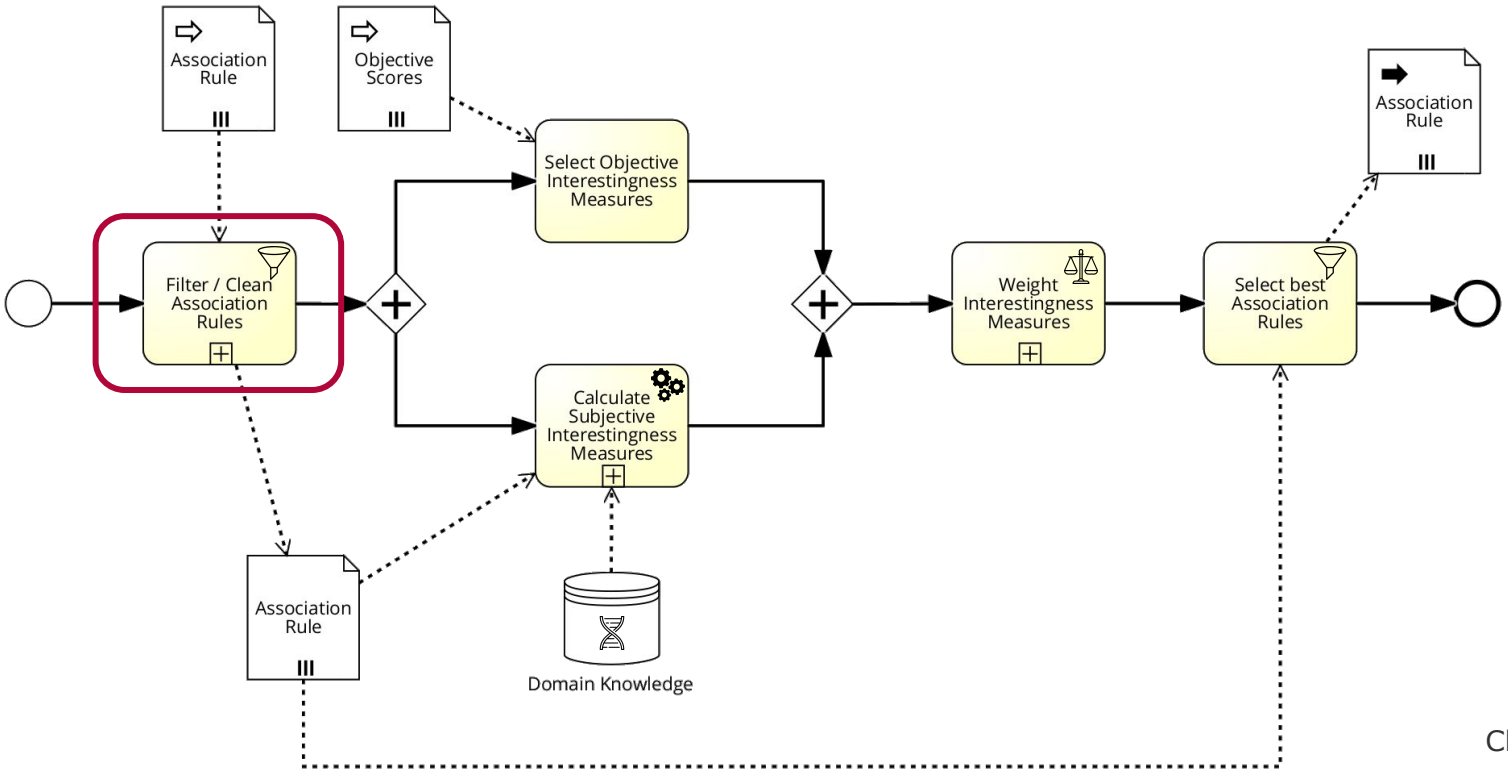
Huge **number of** possible / output **rules**
n frequent items yield $2^n - 2$ association rules
→ which rules are relevant/interesting?

⇒ Select relevant subset of rules

2. Method


Association Rule
Objective Scores
Select Objective Interestingness Measures
Filter / Clean Association Rules
Calculate Subjective Interestingness Measures
Weight Interestingness Measures
Select best Association Rules
Association Rule
Association Rule
Domain Knowledge

2. Method


Association Rule
Objective Scores
Select Objective Interestingness Measures
Filter / Clean Association Rules
Calculate Subjective Interestingness Measures
Weight Interestingness Measures
Select best Association Rules
Association Rule
Association Rule
Domain Knowledge

2. Method - Filtering of Association Rules

Filter association rules that can be removed for formal reasons or because of what the researcher wants to find out or use the results for.

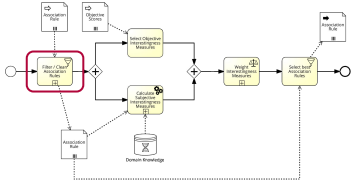
Examples

Item-Filter
Filter association rules containing wanted/unwanted items,
e.g. Gene1— → Gene2—.

Redundancy-Filter
Remove rules that can be deduced from other rules.

If using multiple filters in a row, the order might matter!

2. Method - Min-Max Filter



Min-Max Filter

Idea

The antecedent of a rule should be as small, the consequent as large as possible.

Min-Max$_1$: ensure, that $\frac{|X|}{|Y|} \leq \gamma$ or $|X| \leq max$ hold for rule $X \rightarrow Y$.

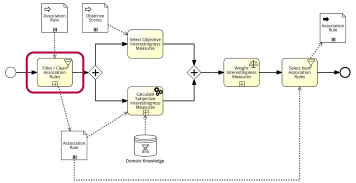
Min-Max$_2$: Given two rules, $r_0 : X_0 \rightarrow Y_0$ and $r_1 : X_1 \rightarrow Y_1$, remove r_0, if $X_1 \subseteq X_0$ and $Y_0 \subseteq Y_1$. If $conf(r_0) = conf(r_1)$, we do not lose information!

Example

r_0: X,Y,Z → C

r_1: X,Y → C

2. Method - Monotonicity



(Strong) Monotonicity

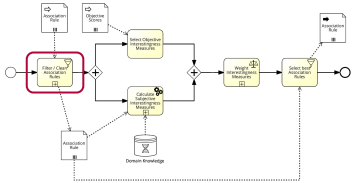
If a rule X → Y holds on a statistically large subset of a dataset D, then X → ¬Y does not hold on any statistically large subset of D.

Example

G2↓ → G4↑
G2↓, G3↑ → G4↓

	Gene1	Gene2	Gene3	Gene4	...
Sample 1	↑	↓		↑	...
Sample 2		↓		↑	...
Sample 3		↓	↑	↓	...
...	...	...	...	...	...

2. Method - Monotonicity Filter

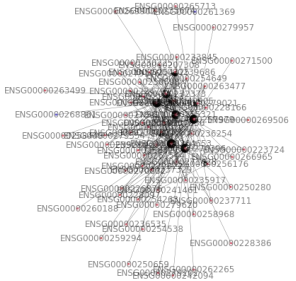


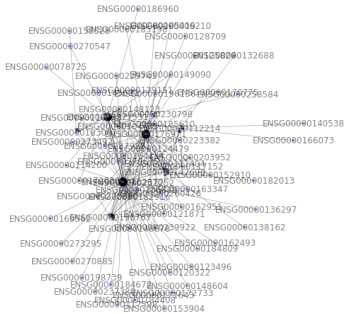
Graph creation

Procedure create_graph(*rules*):
G ← *Graph*()
for *rule* $X \rightarrow Y$ *in rules* **do**
  for $u \in X$ **do**
    for $v \in Y$ **do**
      G.add_edge((u, v))
      G.weights((u, v)).append(*rule.id*)
    end
  end
end
return G

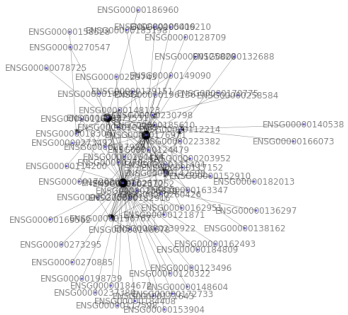
Rule Filtering

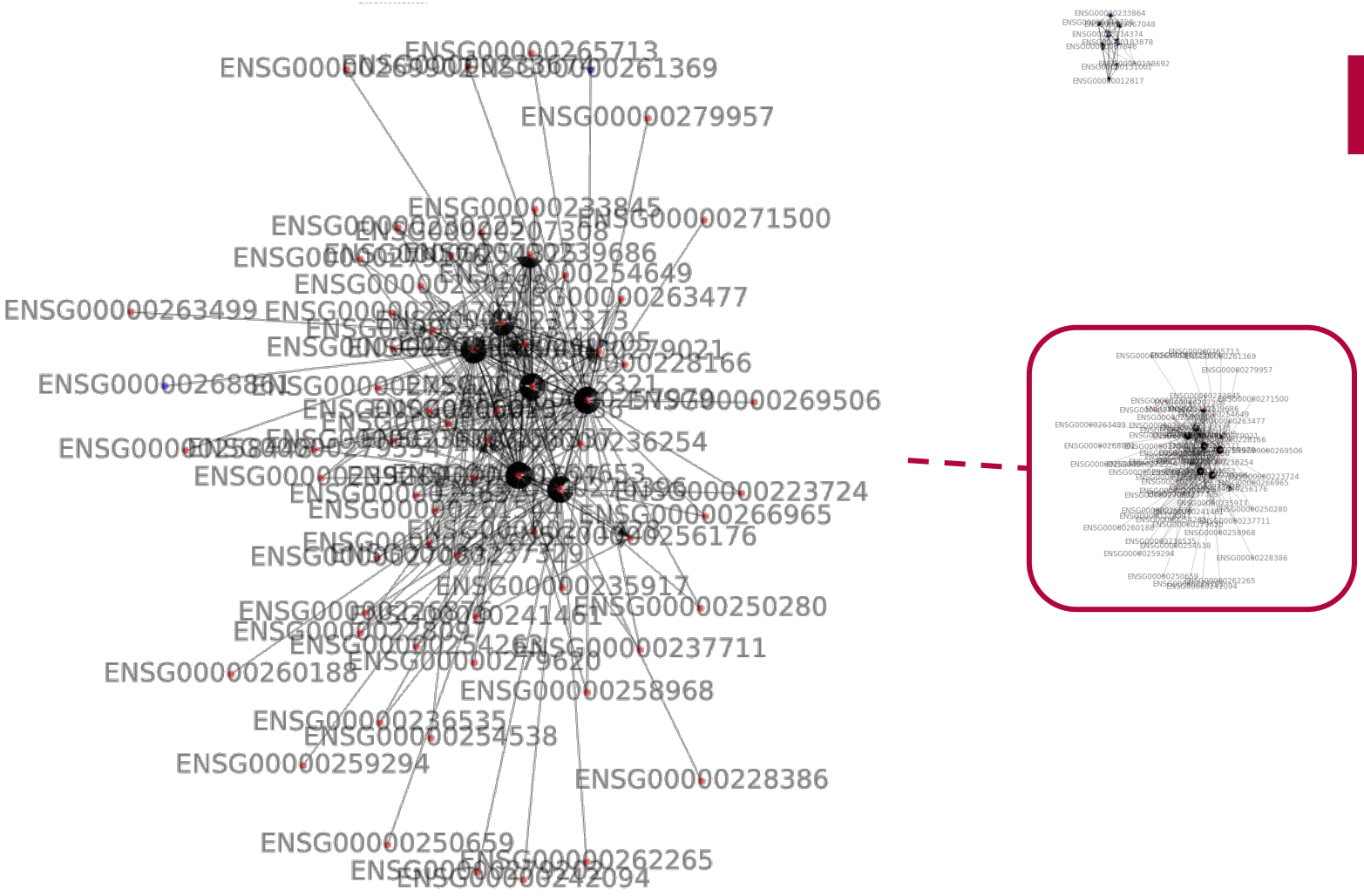
Procedure filter(*rules*):
G ← *create_graph*(rules)
d ← {}
for *node n in G.nodes* **do**
  for *neighbor u* **do**
    for *neighbor v* **do**
      if *opposite(u,v)* **then**
        d ← $d \cup G.weights((n, u))$
        d ← $d \cup G.weights((n, v))$
      end
    end
  end
end
return delete_rules(*rules*,d)

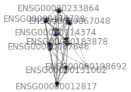


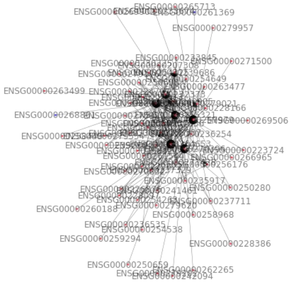


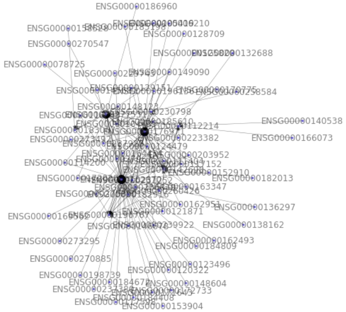
Data: TCGA-LAML_TCGA-GBM

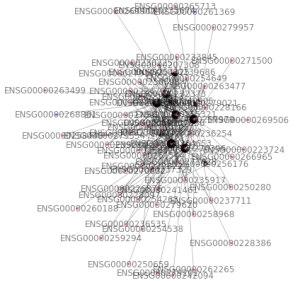



ENSG00000265713
ENSG00000261369
ENSG00000279957
ENSG00000233845
ENSG00000271500
ENSG00000207308
ENSG00000239686
ENSG00000254649
ENSG00000263499
ENSG00000263477
ENSG00000279021
ENSG00000228166
ENSG00000269506
ENSG00000236254
ENSG00000223724
ENSG00000266965
ENSG00000256176
ENSG00000235917
ENSG00000250280
ENSG00000237711
ENSG00000260188
ENSG00000279620
ENSG00000258968
ENSG00000236535
ENSG00000254538
ENSG00000259294
ENSG00000228386
ENSG00000250659
ENSG00000262265
ENSG00000242094

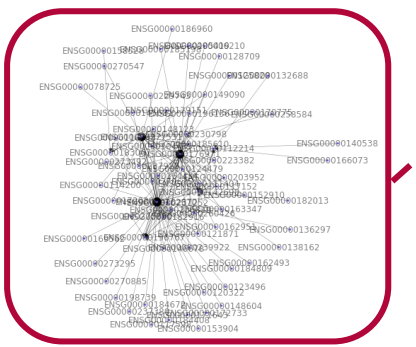




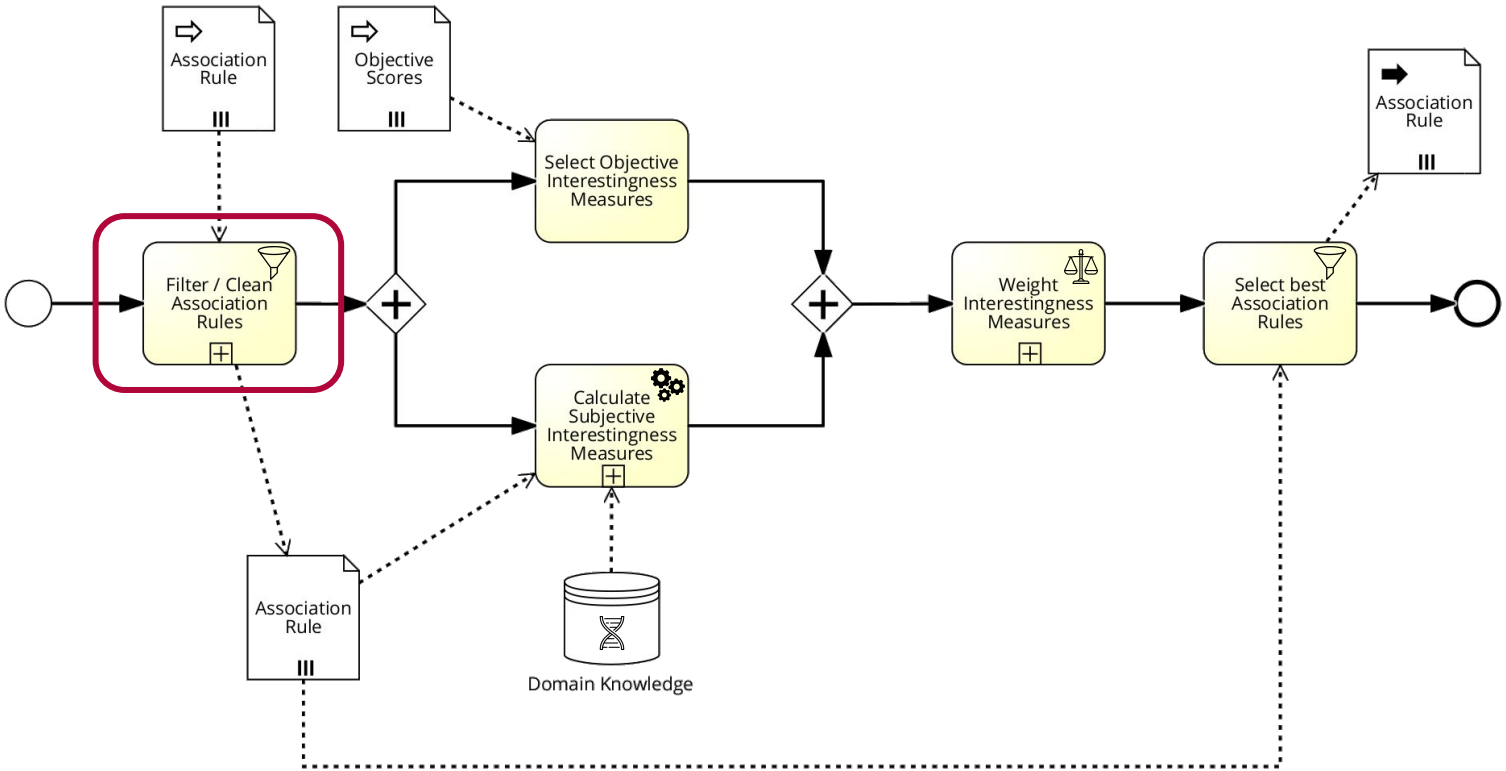




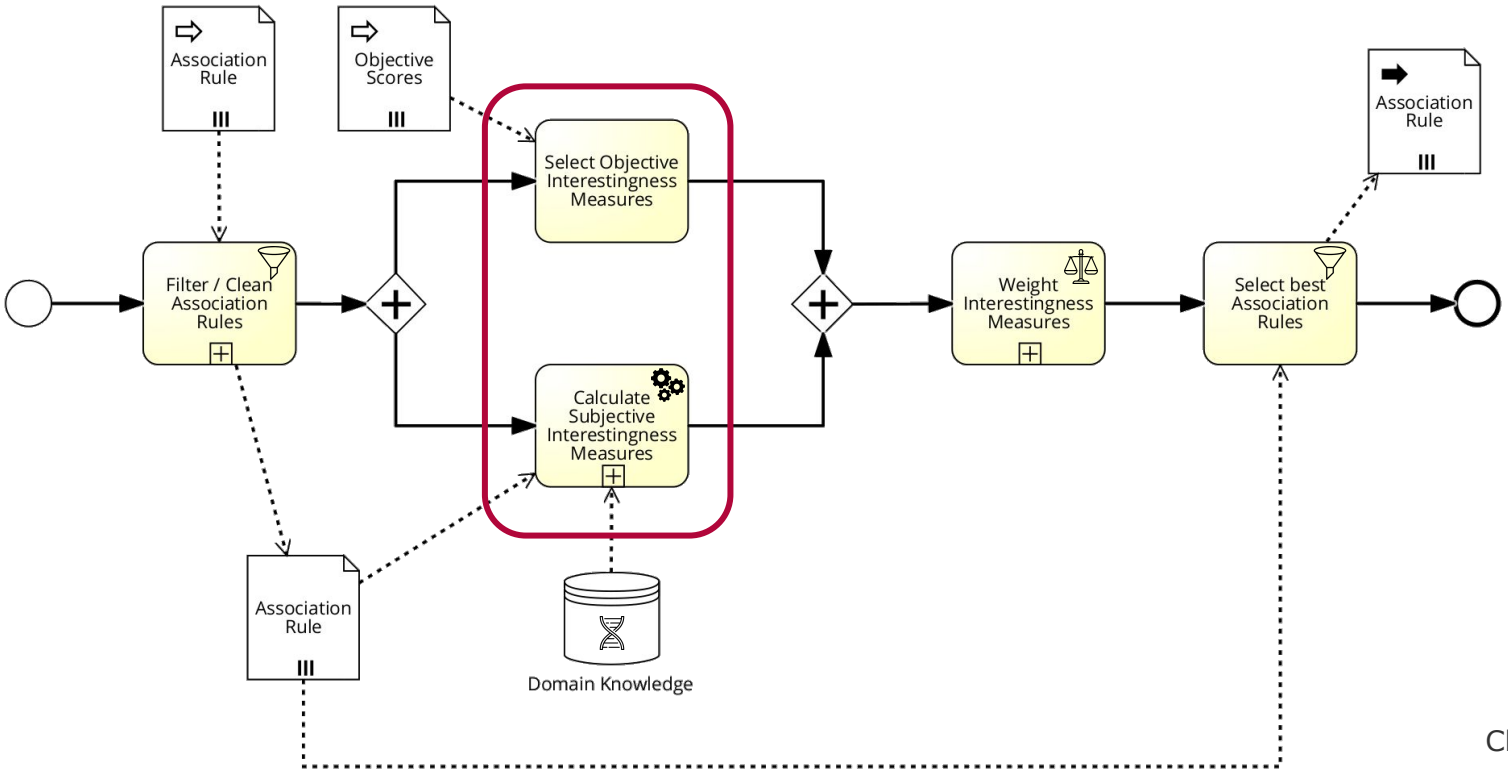
Hasso Plattner Institut


ENSG00000186960
ENSG00000158528
ENSG00000183198
ENSG00000128709
ENSG00000270547
ENSG00000132688
ENSG00000078725
ENSG00000149090
ENSG00000148123
ENSG00000230798
ENSG00000140538
ENSG00000112214
ENSG00000166073
ENSG00000223382
ENSG00000124479
ENSG00000203952
ENSG00000114200
ENSG00000152910
ENSG00000182013
ENSG00000163347
ENSG00000162951
ENSG00000136297
ENSG00000121871
ENSG00000138162
ENSG00000273295
ENSG00000162493
ENSG00000184809
ENSG00000270885
ENSG00000123496
ENSG00000120322
ENSG00000198739
ENSG00000148604
ENSG00000184408
ENSG00000153904

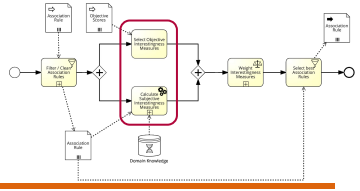
2. Method


Association Rule
Objective Scores
Select Objective Interestingness Measures
Filter / Clean Association Rules
Calculate Subjective Interestingness Measures
Weight Interestingness Measures
Select best Association Rules
Association Rule
Association Rule
Domain Knowledge

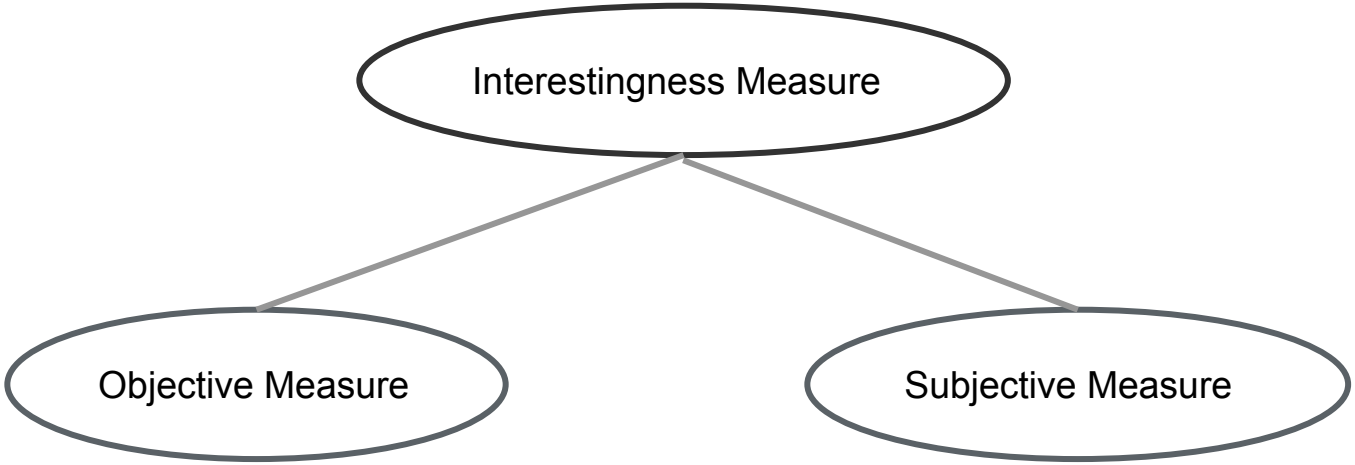
2. Method


Association Rule
Objective Scores
Select Objective Interestingness Measures
Filter / Clean Association Rules
Calculate Subjective Interestingness Measures
Weight Interestingness Measures
Select best Association Rules
Association Rule
Association Rule
Domain Knowledge

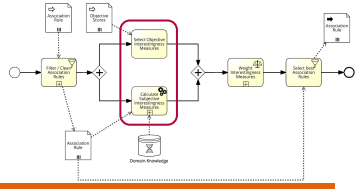
2. Method - Interestingness Measures



Interestingness Measures



2. Method - Objective Interestingness



Interestingness Measures

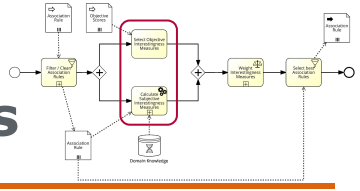
Objective

Depend only on the data and patterns, no knowledge of user or application required. Mostly based on theories in probability, statistics or information theory.

Examples

Support
Confidence
Lift

2. Method - Subjective Interestingness



Interestingness Measures

Subjective

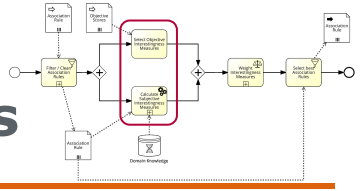
Take into account both the data and the user of these data.

Definition requires domain knowledge (or background knowledge about the data) and its explicit representation.

Examples

Novelty
Unexpectedness
Actionability

2. Method - Subjective Interestingness



External domain knowledge

Genes associated with medical condition:
http://www.disgenet.org/
https://www.targetvalidation.org/

Co-Expression of genes:
http://coxpresdb.jp/
http://www.genefriends.org

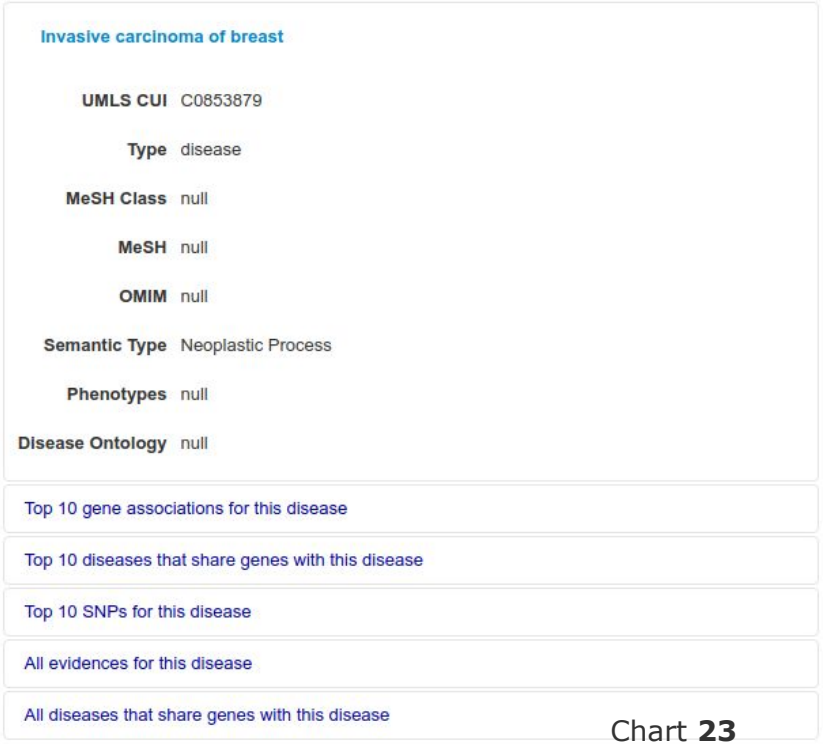
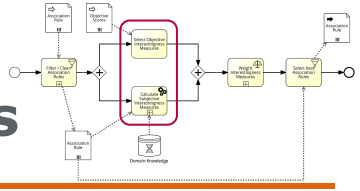


Figure: Disgenet search

2. Method - Subjective Interestingness



First idea:

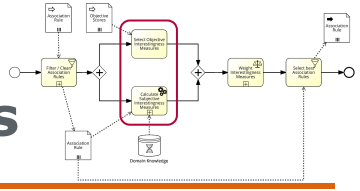
Jaccard-Index: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

Jaccard distance: $d_J(A,B) = 1 - J(A,B)$

Intuition:

Index measures “overlap” / similarity of sets.
If there are lots of genes associated with a medical condition (B),
or a rule (A) is very long, an overlap is more likely,
but the denominator also increases.

2. Method - Subjective Interestingness

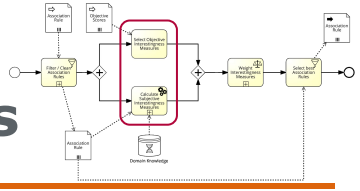


First idea:

For a rule X → Y and a set B of genes associated with a medical condition: $J((X \rightarrow Y), B) = J(X \cup Y, B)$ measures how much a rule corresponds to the domain knowledge.

$d_J((X \rightarrow Y), B)$ measures the novelty of an association rule.

2. Method - Subjective Interestingness



First idea:

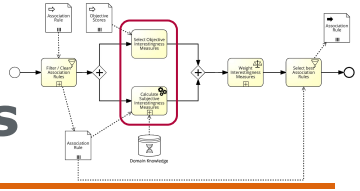
For a rule X → Y and a set B of genes associated with a medical condition:
$J((X \rightarrow Y), B) = J(X \cup Y, B)$ measures how much a rule corresponds to the domain knowledge.

$d_J((X \rightarrow Y), B)$ measures the novelty of an association rule.

Problems:

Scores for different medical conditions hardly comparable
If B is small, $J((X \rightarrow Y), B)$ is almost always 0!

2. Method - Subjective Interestingness



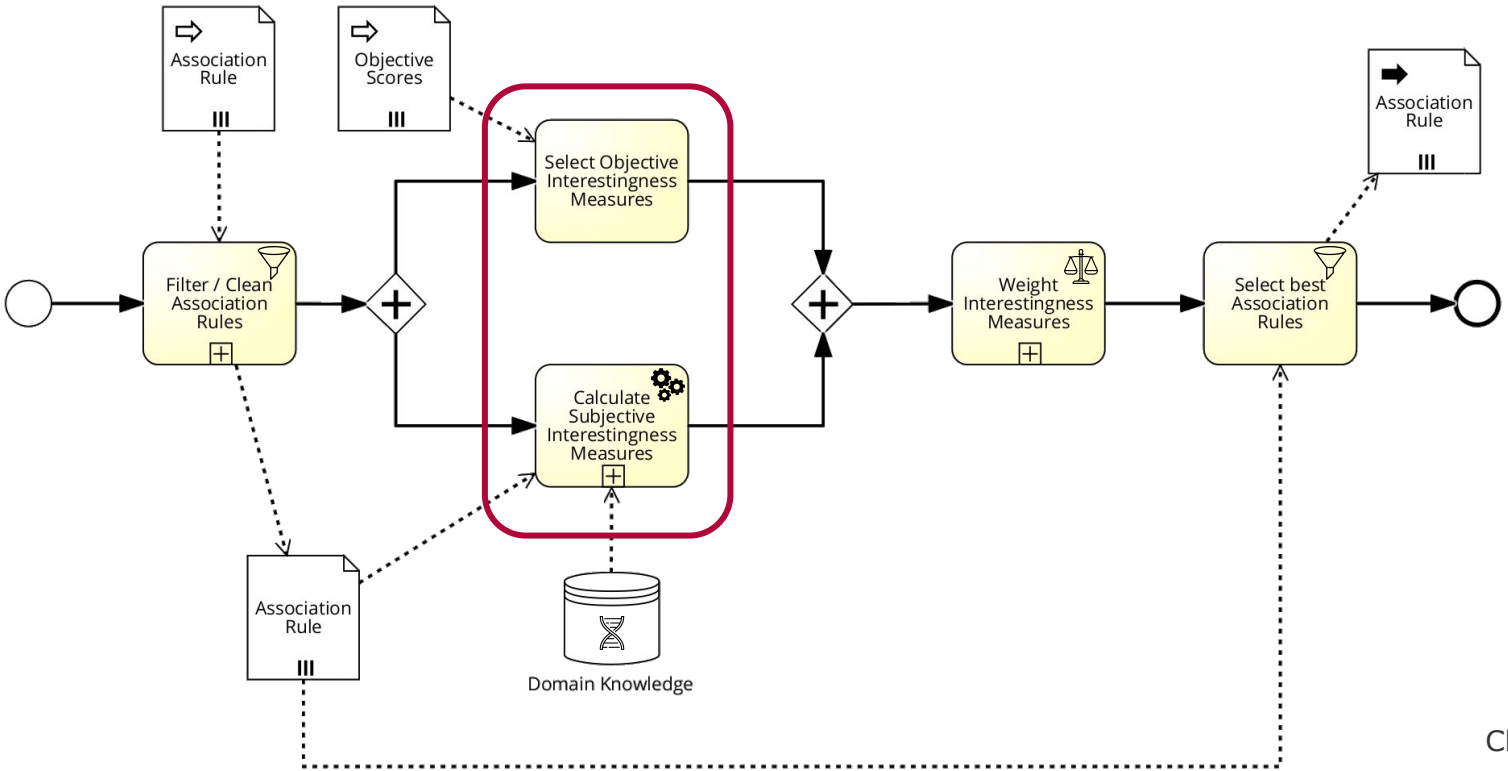
Next idea:

Define a subjective interestingness measure based on gene co-expression.

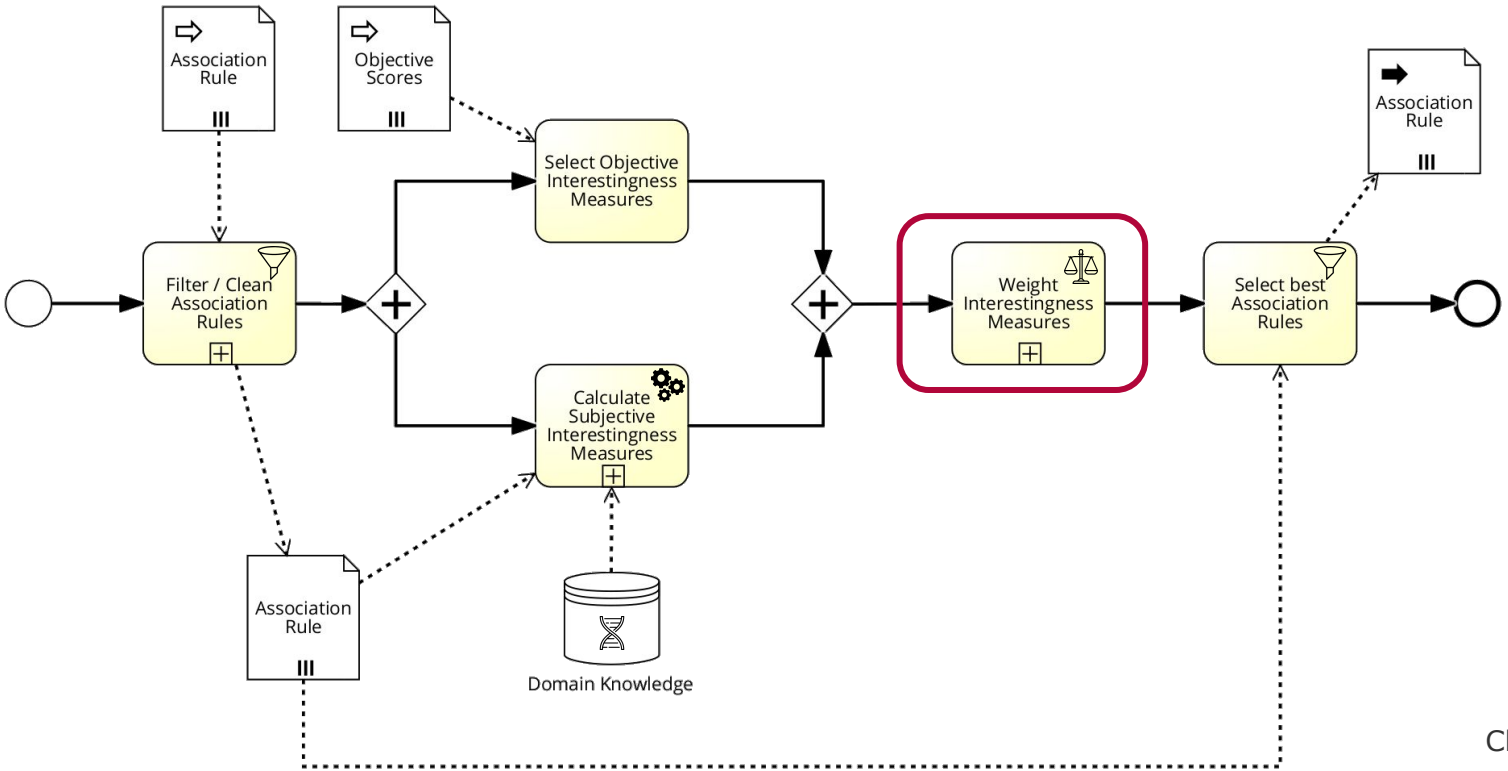
E.g. how strongly do the genes in the antecedent of the association rule correlate with with the genes in the consequent of the association rule?

planned

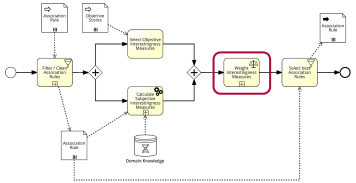
2. Method


Association Rule
Objective Scores
Select Objective Interestingness Measures
Filter / Clean Association Rules
Calculate Subjective Interestingness Measures
Weight Interestingness Measures
Select best Association Rules
Association Rule
Association Rule
Domain Knowledge

2. Method


Association Rule
Objective Scores
Select Objective Interestingness Measures
Filter / Clean Association Rules
Weight Interestingness Measures
Select best Association Rules
Association Rule
Calculate Subjective Interestingness Measures
Association Rule
Domain Knowledge

2. Method - Weighting



Given n interestingness measures in a vector m and an n-dimensional weight vector θ, a combined interesting score can be computed by:

Scalar product

$$M_1(\theta, m) = \langle \theta, m \rangle = \sum_{i=1}^{n} \theta_i m_i$$

Geometric Method

$$M_2(\theta, m) = \prod_{i=1}^{n} m_i^{\theta_i}$$

Setting weights

Value range of m_i
e.g. support $\in \mathbb{N}$, confidence $\in [0, 1]$

Effect of weights in geometric method

$\theta > 1$ increases weight of $m_i > 1$,
but decreases weight of $m_i < 1$

3. Experiments

Setup

Data

RNAseq data from **T**he **C**ancer **G**enome **A**tlas (TCGA)
DSI: TCGA-BRCA_TCGA-PAAD (307 samples; selected 2.7k genes by variance)
DSII: TCGA-LAML_TCGA-GBM (1,28k samples; selected 2.5k genes by variance)

All data sets were discretized by threshold.

Association Rule Miner

Weka 3.9 Apriori Associator

Data	#rules	description
DSI	1000	Breast & Pancreas Carcinomas
DSI^b	4000	*balanced samples*
DSI^p	4000	Pancreas Carcinoma only
DSII	2000	Leukemia & Glioblastoma
$DSII^g$	2000	Glioblastoma samples only

3. Experiments

Experiment #1

Number of filtered rules

1.1 Min-Max$_1$ filter

Data	γ	max	% filtered
DSI	1	-	91.7
DSI	1	1	98.3
DSI	2	-	52.8
DSI	2	2	71.0
DSI	3	-	21.2
DSI	3	3	25.8

Table 1: Min-Max$_1$ Filtering Results

Data	γ	max	% filtered
DSII	1	-	94.3
DSII	1	1	96.9
DSII	2	-	76.9
DSII	2	2	78.3
DSII	3	-	19.5
DSII	3	3	19.5

Table 2: Min-Max$_1$ Filtering Results

3. Experiments

Experiment #1

Number of filtered rules

1.2 Min-Max$_2$ filter

Data	gentle	% filtered
DSI	no	77.2
DSI	yes	49.2
DSI^b	no	78.1
DSI^b	yes	71.5
DSI^p	yes	81.9
DSII	no	45.0
DSII	yes	23.4

Table 3: Min-Max$_2$ Filtering Results

1.3 Monotonicity filter

Data	% filtered
DSI	0.0
DSI^b	0.0
DSI^p	29.4
DSII	0.0
DSII^g	0.0

Table 4: Monotonicity Filtering Results

3. Experiments

Experiment #2 (planned)

Compute average change in rank for association rules after weighting with different subjective interestingness measures.

Experiment #3 (planned)

Evaluate new subjective novelty measure(s) by holding back domain knowledge and see, whether it is discovered.

Experiment #4 (planned)

Retrieve gene network from additional external domain knowledge source and compare to (weighted) association rules / generated graph.

...

4. Conclusion

Filtering techniques are powerful to reduce number of association rules.

Assessing the biological relevance of rules by using subjective interestingness measures proved to be challenging at first try.

Discussion

How to improve the first subjective interestingness measure based on the Jaccard index?

What other subjective interestingness measures are conceivable?

Other ideas regarding experiments and their evaluation?

Association Rule Mining on RNAseq Data

Trends in Bioinformatics

Lukas Ehrig

Jan 23th, 2019

Appendix: Image Sources

Title/Final Slide:
Boehm Konstruktion
http://www.boehm-konstruktion.com/referenzen/hasso-plattner-high-tech-park/

[2] Central Dogma of Molecular Biology:
Khan Academy
https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/intro-to-gene-expression-central-dogma

The BPMN model of the method was made using Signavio Academic Initiative:
https://academic.signavio.com/

Appendix: Gene Expression Data

	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971
TCGA-AB-2841-03B-01T-0760-13	2.76593712005648	1.53324829122169	9.44768726226856	9.88963899009968	9.01877187872937	10.0044407998509	5.69054493055702
TCGA-AB-2818-03A-01T-0734-13	3.58782825427584	1.53324829122169	9.55352739900002	10.0638203943037	9.29156522722249	13.9702018104788	7.95058193099284
TCGA-AB-2976-03A-01T-0734-13	5.56652701424824	2.36734789254045	9.85781380064584	10.1265586746988	9.89295472452555	11.1659358903234	10.2161491140698
TCGA-AB-2867-03A-01T-0734-13	4.4918360998841	1.53324829122169	10.0061015845077	9.45892520299894	9.46391961321915	12.1901941421212	5.59465912838687
TCGA-AB-2839-03A-01T-0734-13	3.40824045628991	1.53324829122169	9.62551673128297	10.3152334019153	9.54101744165605	10.1323934935717	5.07141926665851
TCGA-AB-2881-03A-01T-0735-13	3.42559579476776	1.53324829122169	9.74891883105389	9.91097948514503	9.56450157736863	12.7930569152652	4.618680063385
TCGA-AB-2930-03A-01T-0740-13	2.27803665949948	1.53324829122169	9.6816982615742	9.71862882984999	9.64947557455084	12.0464061583375	6.01813583832155
TCGA-AB-2805-03A-01T-0734-13	1.53324829122169	2.34250535063557	9.93508308009364	9.96642727753544	9.31678991757321	13.875243652276	5.81045856465353
TCGA-AB-2996-03A-01T-0735-13	6.05839766222684	1.53324829122169	9.55610632482587	10.1691079834718	9.98138653010994	12.5302776175638	9.2085392999297
TCGA-AB-2919-03A-01T-0740-13	3.46760139549888	1.53324829122169	9.29621107542366	9.77898483303739	9.96210185508133	8.61696651298998	10.6811598992821
TCGA-AB-2835-03A-01T-0736-13	3.377277044747	1.53324829122169	9.94358747142781	9.09156510537673	9.38790768053172	14.4233187250113	6.60952893809161
TCGA-AB-3012-03A-01T-0736-13	4.79182824926881	1.53324829122169	9.58651623209581	9.47610029001618	8.91034347700386	6.54603396338371	8.19140587613801
TCGA-AB-2842-03A-01T-0734-13	2.3417952587717	1.53324829122169	9.90045946764349	9.99847440855063	9.89905015159871	13.083077404624	6.91285435764463
TCGA-AB-2871-03A-01T-0735-13	4.86264603376552	1.53324829122169	9.11316315950706	10.1237732844438	9.44379726363504	11.0480577028576	7.91698643471225
TCGA-AB-2938-03A-01T-0736-13	7.76952703554482	3.08094347413991	9.11453456835315	9.59464464736721	9.5739925090916	10.6037544757665	12.8757119981577
TCGA-AB-2872-03A-01T-0735-13	4.52927203623542	1.53324829122169	9.20454742690193	9.97858962193787	9.73547163360154	12.3742662794351	7.34163819942233
TCGA-AB-2915-03A-01T-0740-13	3.12189696041535	1.53324829122169	8.80098454848967	10.025719893437	9.27747891349068	13.5059131187465	9.17470944915273
TCGA-AB-2955-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.58945310291011	10.7579572497123	10.5387178077725	9.87641595250127	5.4787977515495
TCGA-AB-2943-03A-01T-0740-13	5.22960414611959	1.53324829122169	9.70534280328187	10.7978009292204	10.2219842837729	11.5693070725036	8.42607343333665
TCGA-AB-2944-03A-01T-0740-13	6.33990221174568	1.53324829122169	9.83248320662875	10.3914671930852	9.6338330041123	10.6956349170102	7.67048168940104
TCGA-AB-3007-03A-01T-0736-13	2.39574027632512	1.53324829122169	9.46332775511307	9.72548645853732	9.29976156052294	7.97681514130557	7.73120968977164
TCGA-AB-2918-03A-01T-0740-13	4.02180530189553	2.33375756783456	9.76440494885102	9.93039644697546	9.70243132635133	11.7858714722287	6.45899654419247
TCGA-AB-2882-03A-01T-0740-13	5.71438877422215	1.53324829122169	9.39106288036326	9.99797579780632	8.97343332402397	12.2208333908276	9.71892498902613
TCGA-AB-2914-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.74888425462813	10.0583377042729	9.71350215738624	11.6649534697277	5.64142817614806
TCGA-AB-2912-03A-01T-0734-13	3.3499271308279	1.53324829122169	10.1816015557517	9.98719613593839	10.2389083180146	12.3412140645443	7.59980012230915

Appendix: Gene Expression Data

	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971
TCGA-AB-2841-03B-01T-0760-13	2.76593712005648	1.53324829122169	9.44768726226856	9.88963899009968	9.01877187872937	10.0044407998509	5.69054493055702
TCGA-AB-2818-03A-01T-0734-13	3.58782825427584	1.53324829122169	9.55352739900002	10.0638203943037	9.29156522722249	13.9702018104788	7.95058193099284
TCGA-AB-2976-03A-01T-0734-13	5.56652701424824	2.36734789254045	9.85781380064584	10.1265586746988	9.89295472452555	11.1659358903234	10.2161491140698
TCGA-AB-2867-03A-01T-0734-13	4.4918360998841	1.53324829122169	10.0061015845077	9.45892520299894	9.46391961321915	12.1901941421212	5.59465912838687
TCGA-AB-2839-03A-01T-0734-13	3.40824045628991	1.53324829122169	9.62551673128297	10.3152334019153	9.54101744165605	10.1323934935717	5.07141926665851
TCGA-AB-2881-03A-01T-0735-13	3.42559579476776	1.53324829122169	9.74891883105389	9.91097948514503	9.56450157736863	12.7930569152652	4.618680063385
TCGA-AB-2930-03A-01T-0740-13	2.27803665949948	1.53324829122169	9.6816982615742	9.71862882984999	9.64947557455084	12.0464061583375	6.01813583832155
TCGA-AB-2805-03A-01T-0734-13	1.53324829122169	2.34250535063557	9.93508308009364	9.96642727753544	9.31678991757321	13.875243652276	5.81045856465353
TCGA-AB-2996-03A-01T-0735-13	6.05839766222684	1.53324829122169	9.55610632482587	10.1691079834718	9.98138653010994	12.5302776175638	9.2085392999297
TCGA-AB-2919-03A-01T-0740-13	3.46760139549888	1.53324829122169	9.29621107542366	9.77898483303739	9.96210185508133	8.61696651298998	10.6811598992821
TCGA-AB-2835-03A-01T-0736-13	3.377277044747	1.53324829122169	9.94358747142781	9.09156510537673	9.38790768053172	14.4233187250113	6.60952893809161
TCGA-AB-3012-03A-01T-0736-13	4.79182824926881	1.53324829122169	9.58651623209581	9.47610029001618	8.91034347700386	6.54603396338371	8.19140587613801
TCGA-AB-2842-03A-01T-0734-13	2.3417952587717	1.53324829122169	9.90045946764349	9.99847440855063	9.89905015159871	13.083077404624	6.91285435764463
TCGA-AB-2871-03A-01T-0735-13	4.86264603376552	1.53324829122169	9.11316315950706	10.1237732844438	9.44379726363504	11.0480577028576	7.91698643471225
TCGA-AB-2938-03A-01T-0736-13	7.76952703554482	3.08094347413991	9.11453456835315	9.59464464736721	9.5739925090916	10.6037544757665	12.8757119981577
TCGA-AB-2872-03A-01T-0735-13	4.52927203623542	1.53324829122169	9.20454742690193	9.97858962193787	9.73547163360154	12.3742662794351	7.34163819942233
TCGA-AB-2915-03A-01T-0740-13	3.12189696041535	1.53324829122169	8.80098454848967	10.025719893437	9.27747891349068	13.5059131187465	9.17470944915273
TCGA-AB-2955-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.58945310291011	10.7579572497123	10.5387178077725	9.87641595250127	5.4787977515495
TCGA-AB-2943-03A-01T-0740-13	5.22960414611959	1.53324829122169	9.70534280328187	10.7978009292204	10.2219842837729	11.5693070725036	8.42607343333665
TCGA-AB-2944-03A-01T-0740-13	6.33990221174568	1.53324829122169	9.83248320662875	10.3914671930852	9.6338330041123	10.6956349170102	7.67048168940104
TCGA-AB-3007-03A-01T-0736-13	2.39574027632512	1.53324829122169	9.46332775511307	9.72548645853732	9.29976156052294	7.97681514130557	7.73120968977164
TCGA-AB-2918-03A-01T-0740-13	4.02180530189553	2.33375756783456	9.76440494885102	9.93039644697546	9.70243132635133	11.7858714722287	6.45899654419247
TCGA-AB-2882-03A-01T-0740-13	5.71438877422215	1.53324829122169	9.39106288036326	9.99797579780632	8.97343332402397	12.2208333908276	9.71892498902613
TCGA-AB-2914-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.74888425462813	10.0583377042729	9.71350215738624	11.6649534697277	5.64142817614806
TCGA-AB-2912-03A-01T-0734-13	3.3499271308279	1.53324829122169	10.1816015557517	9.98719613593839	10.2389083180146	12.3412140645443	7.59980012230915

Gene

Appendix: Gene Expression Data

	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971
TCGA-AB-2841-03B-01T-0760-13	2.76593712005648	1.53324829122169	9.44768726226856	9.88963899009968	9.01877187872937	10.0044407998509	5.69054493055702
TCGA-AB-2818-03A-01T-0734-13	3.58782825427584	1.53324829122169	9.55352739900002	10.0638203943037	9.29156522722249	13.9702018104788	7.95058193099284
TCGA-AB-2976-03A-01T-0734-13	5.56652701424824	2.36734789254045	9.85781380064584	10.1265586746988	9.89295472452555	11.1659358903234	10.2161491140698
TCGA-AB-2867-03A-01T-0734-13	4.4918360998841	1.53324829122169	10.0061015845077	9.45892520299894	9.46391961321915	12.1901941421212	5.59465912838687
TCGA-AB-2839-03A-01T-0734-13	3.40824045628991	1.53324829122169	9.62551673128297	10.3152334019153	9.54101744165605	10.1323934935717	5.07141926665851
TCGA-AB-2881-03A-01T-0735-13	3.42559579476776	1.53324829122169	9.74891883105389	9.91097948514503	9.56450157736863	12.7930569152652	4.618680063385
TCGA-AB-2930-03A-01T-0740-13	2.27803665949948	1.53324829122169	9.6816982615742	9.71862882984999	9.64947557455084	12.0464061583375	6.01813583832155
TCGA-AB-2805-03A-01T-0734-13	1.53324829122169	2.34250535063557	9.93508308009364	9.96642727753544	9.31678991757321	13.875243652276	5.81045856465353
TCGA-AB-2996-03A-01T-0735-13	6.05839766222684	1.53324829122169	9.55610632482587	10.1691079834718	9.98138653010994	12.5302776175638	9.2085392999297
TCGA-AB-2919-03A-01T-0740-13	3.46760139549888	1.53324829122169	9.29621107542366	9.77898483303739	9.96210185508133	8.61696651298998	10.6811598992821
TCGA-AB-2835-03A-01T-0736-13	3.377277044747	1.53324829122169	9.94358747142781	9.09156510537673	9.38790768053172	14.4233187250113	6.60952893809161
TCGA-AB-3012-03A-01T-0736-13	4.79182824926881	1.53324829122169	9.58651623209581	9.47610029001618	8.91034347700386	6.54603396338371	8.19140587613801
TCGA-AB-2842-03A-01T-0734-13	2.3417952587717	1.53324829122169	9.90045946764349	9.99847440855063	9.89905015159871	13.083077404624	6.91285435764463
TCGA-AB-2871-03A-01T-0735-13	4.86264603376552	1.53324829122169	9.11316315950706	10.1237732844438	9.44379726363504	11.0480577028576	7.91698643471225
TCGA-AB-2938-03A-01T-0736-13	7.76952703554482	3.08094347413991	9.11453456835315	9.59464464736721	9.5739925090916	10.6037544757665	12.8757119981577
TCGA-AB-2872-03A-01T-0735-13	4.52927203623542	1.53324829122169	9.20454742690193	9.97858962193787	9.73547163360154	12.3742662794351	7.34163819942233
TCGA-AB-2915-03A-01T-0740-13	3.12189696041535	1.53324829122169	8.80098454848967	10.025719893437	9.27747891349068	13.5059131187465	9.17470944915273
TCGA-AB-2955-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.58945310291011	10.7579572497123	10.5387178077725	9.87641595250127	5.4787977515495
TCGA-AB-2943-03A-01T-0740-13	5.22960414611959	1.53324829122169	9.70534280328187	10.7978009292204	10.2219842837729	11.5693070725036	8.42607343333665
TCGA-AB-2944-03A-01T-0740-13	6.33990221174568	1.53324829122169	9.83248320662875	10.3914671930852	9.6338330041123	10.6956349170102	7.67048168940104
TCGA-AB-3007-03A-01T-0736-13	2.39574027632512	1.53324829122169	9.46332775511307	9.72548645853732	9.29976156052294	7.97681514130557	7.73120968977164
TCGA-AB-2918-03A-01T-0740-13	4.02180530189553	2.33375756783456	9.76440494885102	9.93039644697546	9.70243132635133	11.7858714722287	6.45899654419247
TCGA-AB-2882-03A-01T-0740-13	5.71438877422215	1.53324829122169	9.39106288036326	9.99797579780632	8.97343332402397	12.2208333908276	9.71892498902613
TCGA-AB-2914-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.74888425462813	10.0583377042729	9.71350215738624	11.6649534697277	5.64142817614806
TCGA-AB-2912-03A-01T-0734-13	3.3499271308279	1.53324829122169	10.1816015557517	9.98719613593839	10.2389083180146	12.3412140645443	7.59980012230915

Sample

Appendix: Gene Expression Data

	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971
TCGA-AB-2841-03B-01T-0760-13	2.76593712005648	1.53324829122169	9.44768726226856	9.88963899009968	9.01877187872937	10.0044407998509	5.69054493055702
TCGA-AB-2818-03A-01T-0734-13	3.58782825427584	1.53324829122169	9.55352739900002	10.0638203943037	9.29156522722249	13.9702018104788	7.95058193099284
TCGA-AB-2976-03A-01T-0734-13	5.56652701424824	2.36734789254045	9.85781380064584	10.1265586746988	9.89295472452555	11.1659358903234	10.2161491140698
TCGA-AB-2867-03A-01T-0734-13	4.4918360998841	1.53324829122169	10.0061015845077	9.45892520299894	9.46391961321915	12.1901941421212	5.59465912838687
TCGA-AB-2839-03A-01T-0734-13	3.40824045628991	1.53324829122169	9.62551673128297	10.3152334019153	9.54101744165605	10.1323934935717	5.07141926665851
TCGA-AB-2881-03A-01T-0735-13	3.42559579476776	1.53324829122169	9.74891883105389	9.91097948514503	9.56450157736863	12.7930569152652	4.618680063385
TCGA-AB-2930-03A-01T-0740-13	2.27803665949948	1.53324829122169	9.6816982615742	9.71862882984999	9.64947557455084	12.0464061583375	6.01813583832155
TCGA-AB-2805-03A-01T-0734-13	1.53324829122169	2.34250535063557	9.93508308009364	9.96642727753544	9.31678991757321	13.875243652276	5.81045856465353
TCGA-AB-2996-03A-01T-0735-13	6.05839766222684	1.53324829122169	9.55610632482587	10.1691079834718	9.98138653010994	12.5302776175638	9.2085392999297
TCGA-AB-2919-03A-01T-0740-13	3.46760139549	1.53324829122169	9.296211075423	9.7899483303739	9.96210185508133	8.61696651298998	10.6811598992821
TCGA-AB-2835-03A-01T-0736-13	3.37727044747	1.53	9.9	9.9	9.38790768053172	14.4233187250113	6.60952893809161
TCGA-AB-3012-03A-01T-0736-13	4.79182824926	2.53	9.5	9.7	8.91034347700386	6.54603396338371	8.19140587613801
TCGA-AB-2842-03A-01T-0734-13	2.3417952587717	1.53324829122169	9.90045946764349	9.99847440855063	8.89905015159871	13.083077404624	6.91285435764463
TCGA-AB-2871-03A-01T-0735-13	4.86264603376552	1.53324829122169	9.11316315950706	10.1237732844438	9.44379726363504	11.0480577028576	7.91698643471225
TCGA-AB-2938-03A-01T-0736-13	7.76952703554482	3.08094347413991	9.11453456835315	9.59464464736721	9.5739925090916	10.6037544757665	12.8757119981577
TCGA-AB-2872-03A-01T-0735-13	4.52927203623542	1.53324829122169	9.20454742690193	9.97858962193787	9.73547163360154	12.3742662794351	7.34163819942233
TCGA-AB-2915-03A-01T-0740-13	3.12189696041535	1.53324829122169	8.80098454848967	10.025719893437	9.27747891349068	13.5059131187465	9.17470944915273
TCGA-AB-2955-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.58945310291011	10.7579572497123	10.5387178077725	9.87641595250127	5.4787977515495
TCGA-AB-2943-03A-01T-0740-13	5.22960414611959	1.53324829122169	9.70534280328187	10.7978009292204	10.2219842837729	11.5693070725036	8.42607343333665
TCGA-AB-2944-03A-01T-0740-13	6.33990221174568	1.53324829122169	9.83248320662875	10.3914671930852	9.6338330041123	10.6956349170102	7.67048168940104
TCGA-AB-3007-03A-01T-0736-13	2.39574027632512	1.53324829122169	9.46332775511307	9.72548645853732	9.29976156052294	7.97681514130557	7.73120968977164
TCGA-AB-2918-03A-01T-0740-13	4.02180530189553	2.33375756783456	9.76440494885102	9.93039644697546	9.70243132635133	11.7858714722287	6.45899654419247
TCGA-AB-2882-03A-01T-0740-13	5.71438877422215	1.53324829122169	9.39106288036326	9.99797579780632	8.97343332402397	12.2208333908276	9.71892498902613
TCGA-AB-2914-03A-01T-0734-13	1.53324829122169	1.53324829122169	9.74888425462813	10.0583377042729	9.71350215738624	11.6649534697277	5.64142817614806
TCGA-AB-2912-03A-01T-0734-13	3.3499271308279	1.53324829122169	10.1816015557517	9.98719613593839	10.2389083180146	12.3412140645443	7.59980012230915

Discretize

Appendix: Gene Expression Data

	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971
TCGA-AB-2841-03B-01T-0760-13	-1.0	0.0	0.0	0.0	0.0	0.0	-1.0
TCGA-AB-2818-03A-01T-0734-13	-1.0	0.0	0.0	1.0	0.0	1.0	0.0
TCGA-AB-2976-03A-01T-0734-13	0.0	0.0	0.0	1.0	1.0	0.0	0.0
TCGA-AB-2867-03A-01T-0734-13	0.0	0.0	0.0	0.0	0.0	0.0	-1.0
TCGA-AB-2839-03A-01T-0734-13	-1.0	0.0	0.0	1.0	0.0	0.0	-1.0
TCGA-AB-2881-03A-01T-0735-13	-1.0	0.0	0.0	0.0	0.0	1.0	-1.0
TCGA-AB-2930-03A-01T-0740-13	-1.0	0.0	0.0	0.0	0.0	0.0	-1.0
TCGA-AB-2805-03A-01T-0734-13	-1.0	0.0	0.0	1.0	0.0	1.0	-1.0
TCGA-AB-2996-03A-01T-0735-13	0.0	0.0	0.0	1.0	1.0	1.0	0.0
TCGA-AB-2919-03A-01T-0740-13	-1.0	0.0	-1.0	0.0	1.0	0.0	0.0
TCGA-AB-2835-03A-01T-0736-13	-1.0	0.0	0.0	0.0	0.0	1.0	-1.0
TCGA-AB-3012-03A-01T-0736-13	0.0	0.0	0.0	0.0	0.0	-1.0	0.0
TCGA-AB-2842-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	1.0	0.0
TCGA-AB-2871-03A-01T-0735-13	0.0	0.0	-1.0	1.0	0.0	0.0	0.0
TCGA-AB-2938-03A-01T-0736-13	0.0	0.0	-1.0	0.0	0.0	0.0	1.0
TCGA-AB-2872-03A-01T-0735-13	0.0	0.0	-1.0	1.0	1.0	1.0	0.0
TCGA-AB-2915-03A-01T-0740-13	-1.0	0.0	-1.0	1.0	0.0	1.0	0.0
TCGA-AB-2955-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	0.0	-1.0
TCGA-AB-2943-03A-01T-0740-13	0.0	0.0	0.0	1.0	1.0	0.0	0.0
TCGA-AB-2944-03A-01T-0740-13	0.0	0.0	0.0	1.0	0.0	0.0	0.0
TCGA-AB-3007-03A-01T-0736-13	-1.0	0.0	0.0	0.0	0.0	-1.0	0.0
TCGA-AB-2918-03A-01T-0740-13	0.0	0.0	0.0	1.0	1.0	0.0	-1.0
TCGA-AB-2882-03A-01T-0740-13	0.0	0.0	-1.0	1.0	0.0	0.0	0.0
TCGA-AB-2914-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	0.0	-1.0
TCGA-AB-2912-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	1.0	0.0

Appendix: Association Rule Mining

Table 1. Market basket transactions

Transaction ID	*Items Bought*
1	{*Laptop, Printer, Tablet, Headset*}
2	{*Printer, Monitor, Tablet*}
3	{*Laptop, Printer, Tablet, Headset*}
4	{*Laptop, Monitor, Tablet, Headset*}
5	{*Printer, Monitor, Tablet, Headset*}
6	{*Printer, Tablet, Headset*}
7	{*Monitor, Tablet*}
8	{*Laptop, Printer, Monitor*}
9	{*Laptop, Tablet, Headset*}
10	{*Printer, Tablet*}

{Headset,Laptop}

Appendix: Association Rule Mining

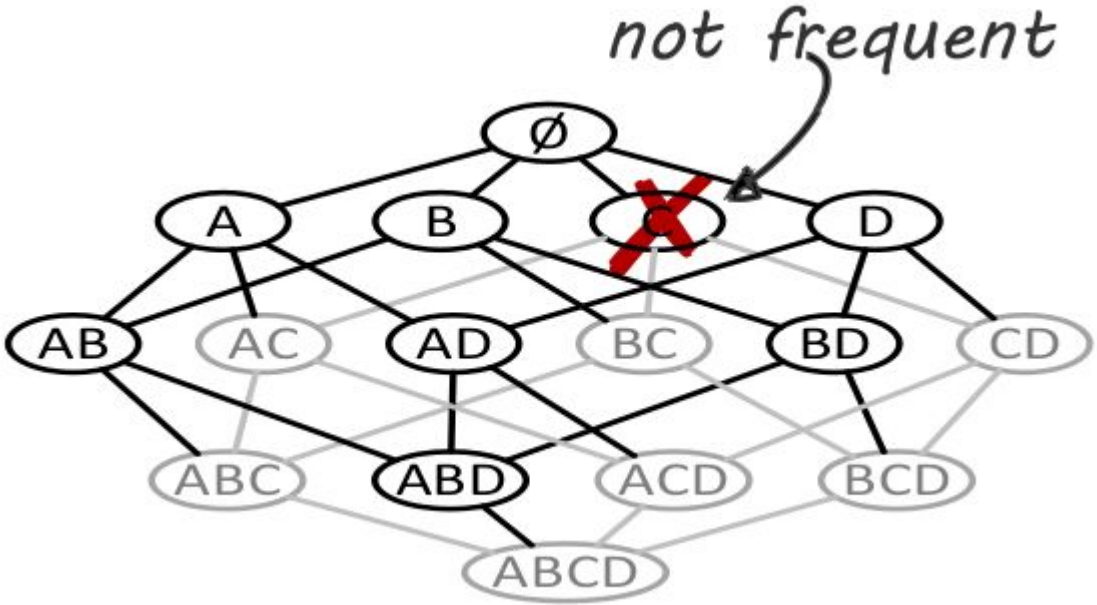
Table 1. Market basket transactions

Transaction ID	*Items Bought*
1	{*Laptop, Printer, Tablet, Headset*}
2	{*Printer, Monitor, Tablet*}
3	{*Laptop, Printer, Tablet, Headset*}
4	{*Laptop, Monitor, Tablet, Headset*}
5	{*Printer, Monitor, Tablet, Headset*}
6	{*Printer, Tablet, Headset*}
7	{*Monitor, Tablet*}
8	{*Laptop, Printer, Monitor*}
9	{*Laptop, Tablet, Headset*}
10	{*Printer, Tablet*}

A: Laptop ⇒ Headset
sup(A) = 4
conf(A) = 0.8

B: Headset ⇒ Laptop
sup(B) = 4
conf(B) = 0.6̄

Appendix: Association Rule Mining



Appendix: Association Rule Mining

	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971
TCGA-AB-2841-03B-01T-0760-13	-1.0	0.0	0.0	0.0	0.0	0.0	-1.0
TCGA-AB-2818-03A-01T-0734-13	-1.0	0.0	0.0	1.0	0.0	1.0	0.0
TCGA-AB-2976-03A-01T-0734-13	0.0	0.0	0.0	1.0	1.0	0.0	0.0
TCGA-AB-2867-03A-01T-0734-13	0.0	0.0	0.0	0.0	0.0	0.0	-1.0
TCGA-AB-2839-03A-01T-0734-13	-1.0	0.0	0.0	1.0	0.0	0.0	-1.0
TCGA-AB-2881-03A-01T-0735-13	-1.0	0.0	0.0	0.0	0.0	1.0	-1.0
TCGA-AB-2930-03A-01T-0740-13	-1.0	0.0	0.0	0.0	0.0	0.0	-1.0
TCGA-AB-2805-03A-01T-0734-13	-1.0	0.0	0.0	1.0	0.0	1.0	-1.0
TCGA-AB-2996-03A-01T-0735-13	0.0	0.0	0.0	1.0	1.0	1.0	0.0
TCGA-AB-2919-03A-01T-0740-13	-1.0	0.0	-1.0	0.0	1.0	0.0	0.0
TCGA-AB-2835-03A-01T-0736-13	-1.0	0.0	0.0	0.0	0.0	1.0	-1.0
TCGA-AB-3012-03A-01T-0736-13	0.0	0.0	0.0	0.0	0.0	-1.0	0.0
TCGA-AB-2842-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	1.0	0.0
TCGA-AB-2871-03A-01T-0735-13	0.0	0.0	-1.0	1.0	0.0	0.0	0.0
TCGA-AB-2938-03A-01T-0736-13	0.0	0.0	-1.0	0.0	0.0	0.0	1.0
TCGA-AB-2872-03A-01T-0735-13	0.0	0.0	-1.0	1.0	1.0	1.0	0.0
TCGA-AB-2915-03A-01T-0740-13	-1.0	0.0	-1.0	1.0	0.0	1.0	0.0
TCGA-AB-2955-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	0.0	-1.0
TCGA-AB-2943-03A-01T-0740-13	0.0	0.0	0.0	1.0	1.0	0.0	0.0
TCGA-AB-2944-03A-01T-0740-13	0.0	0.0	0.0	1.0	0.0	0.0	0.0
TCGA-AB-3007-03A-01T-0736-13	-1.0	0.0	0.0	0.0	0.0	-1.0	0.0
TCGA-AB-2918-03A-01T-0740-13	0.0	0.0	0.0	1.0	1.0	0.0	-1.0
TCGA-AB-2882-03A-01T-0740-13	0.0	0.0	-1.0	1.0	0.0	0.0	0.0
TCGA-AB-2914-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	0.0	-1.0
TCGA-AB-2912-03A-01T-0734-13	-1.0	0.0	0.0	1.0	1.0	1.0	0.0

Item: ENSG0..03=-1

Transaction ID

Transaction

Appendix: Redundancy Filter

Redundant Association Rules

An association rule is redundant if its structure and statistical measures can be deduced from another rule.

Different notions of redundancy. Based on Galois-closure according to Zaki (for approximate association rules):

[Mohammed J. Zaki. 2004. Mining Non-Redundant Association Rules. Data Min. Knowl. Discov. 9, 3 (Nov. 2004), 223-248.]

Appendix: Non-redundant Association Rules

Theorem 4.2. *Let* $\mathcal{R} = \{R_1, \ldots, R_n\}$ *be the set of all possible rules that satisfy the following conditions*:

1. $q_i = q$ *for all* $1 \leq i \leq n$ (*i.e., all rules have the same support*).
2. $p_i = p < 1.0$ *for all* $1 \leq i \leq n$ (*i.e., all rules have same confidence*).
3. $I_1 = c_{it}(X_1^i)$, *and* $I_2 = c_{it}(X_1^i \cup X_2^i)$ *for all* $1 \leq i \leq n$.

Let $\mathcal{R}^G = \{R_i \mid \nexists R_j \in \mathcal{R}, R_j \prec R_i\}$, *denote the most general rules in* $\mathcal{R}$. *Then all rules* $R_i \in \mathcal{R}$ *are equivalent to the rule* $I_1 \xrightarrow{q,p} I_2$, *and all rules in* $\mathcal{R} - \mathcal{R}^G$ *are redundant.*

[Mohammed J. Zaki. 2004. Mining Non-Redundant Association Rules. Data Min. Knowl. Discov. 9, 3 (Nov. 2004), 223-248.]

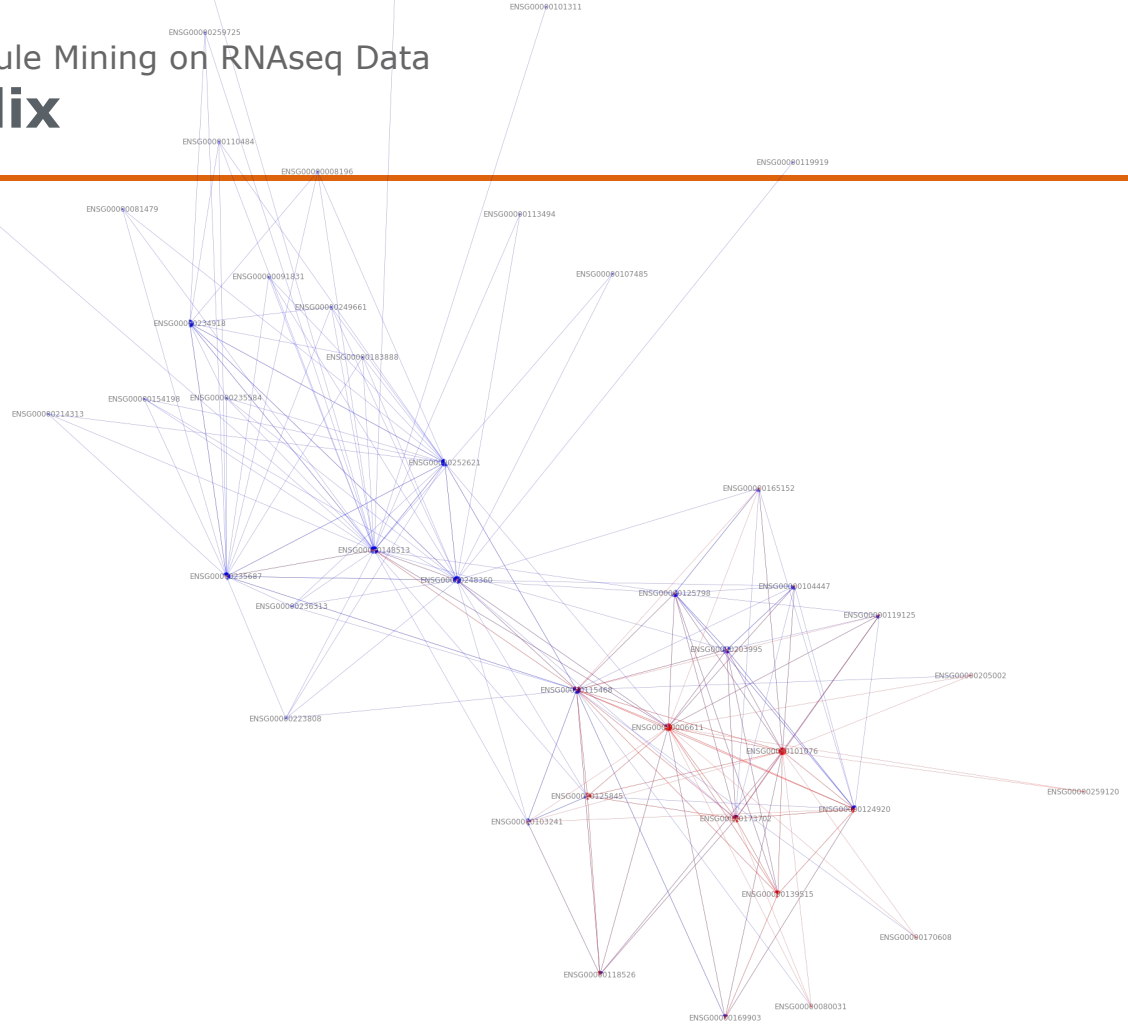
Appendix: Non-redundant Association Rules

Proof: Consider any rule $R_i = X_1^i \xrightarrow{p} X_2^i$. Then the support of the rule is given as $q = |t(X_1^i \cup X_2^i)|$ and its confidence as $p = q/d$, $d = |t(X_1^i)|$. We will show that the $I_1 \longrightarrow I_2$ also has support $|t(I_1 \cup I_2)| = q$ and confidence $\frac{|t(I_1 \cup I_2)|}{|t(I_1)|} = q/d$.

Let's consider the denominator first. We have $|t(I_1)| = |t(c_{it}(X_1^i))| = |t(X_1^i)| = d$. Now consider the numerator. We have $|t(I_1 \cup I_2)| = |t(c_{it}(X_1^i) \cup c_{it}(X_1^i \cup X_2^i))|$. Since $X_1^i \subseteq (X_1^i \cup X_2^i)$, we have, from the property of closure operator, $c_{it}(X_1^i) \subseteq c_{it}(X_1^i \cup X_2^i)$. Thus, $|t(I_1 \cup I_2)| = |t(c_{it}(X_1^i \cup X_2^i))| = |t(X_1^i \cup X_2^i)| = q$. □

[Mohammed J. Zaki. 2004. Mining Non-Redundant Association Rules. Data Min. Knowl. Discov. 9, 3 (Nov. 2004), 223-248.]

Appendix


ENSG00000101311
ENSG00000259725
ENSG00000110484
ENSG00000136944
ENSG00000008196
ENSG00000119919
ENSG00000081479
ENSG00000113494
ENSG00000091831
ENSG00000107485
ENSG00000249661
ENSG00000234918
ENSG00000183888
ENSG00000154198
ENSG00000235584
ENSG00000214313
ENSG00000252621
ENSG00000165152
ENSG00000148513
ENSG00000235687
ENSG00000248360
ENSG00000125798
ENSG00000104447
ENSG00000238513
ENSG00000119125
ENSG00000203995
ENSG00000205002
ENSG00000115468
ENSG00000223808
ENSG00000206611
ENSG00000101976
ENSG00000259120
ENSG00000125845
ENSG00000124920
ENSG00000103241
ENSG00000177707
ENSG00000139515
ENSG00000170608
ENSG00000118526
ENSG00000080031
ENSG00000169903