SEMANTIC BICLUSTERING WITH RULES

František Malinka

RuleML Webinar

27/4/2022

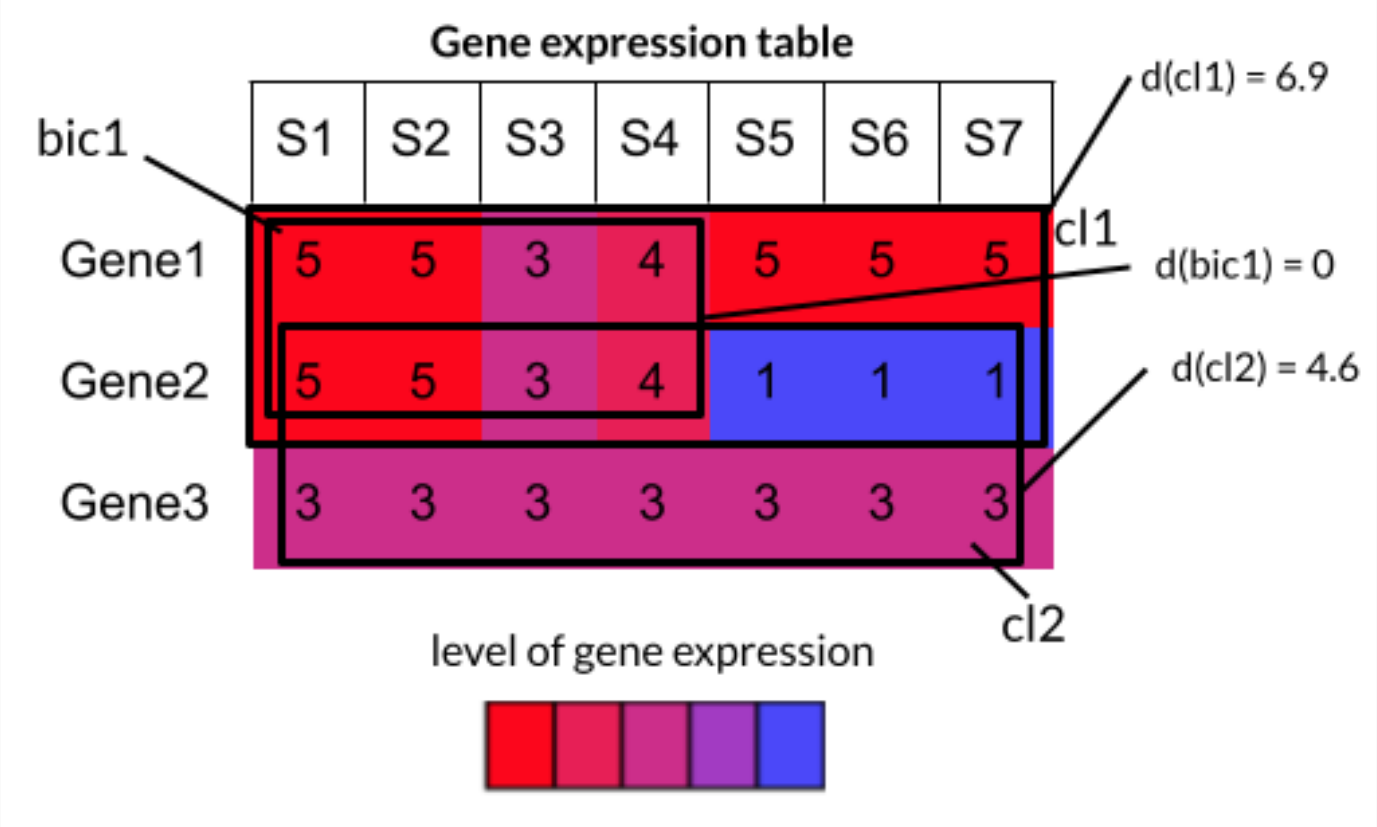
Czech Technical University in Prague

Outline

1. Biclustering - a popular approach
2. Semantic biclustering - idea
3. Two approaches
4. Efficient ontology operator
5. Experiments
6. Conclusions

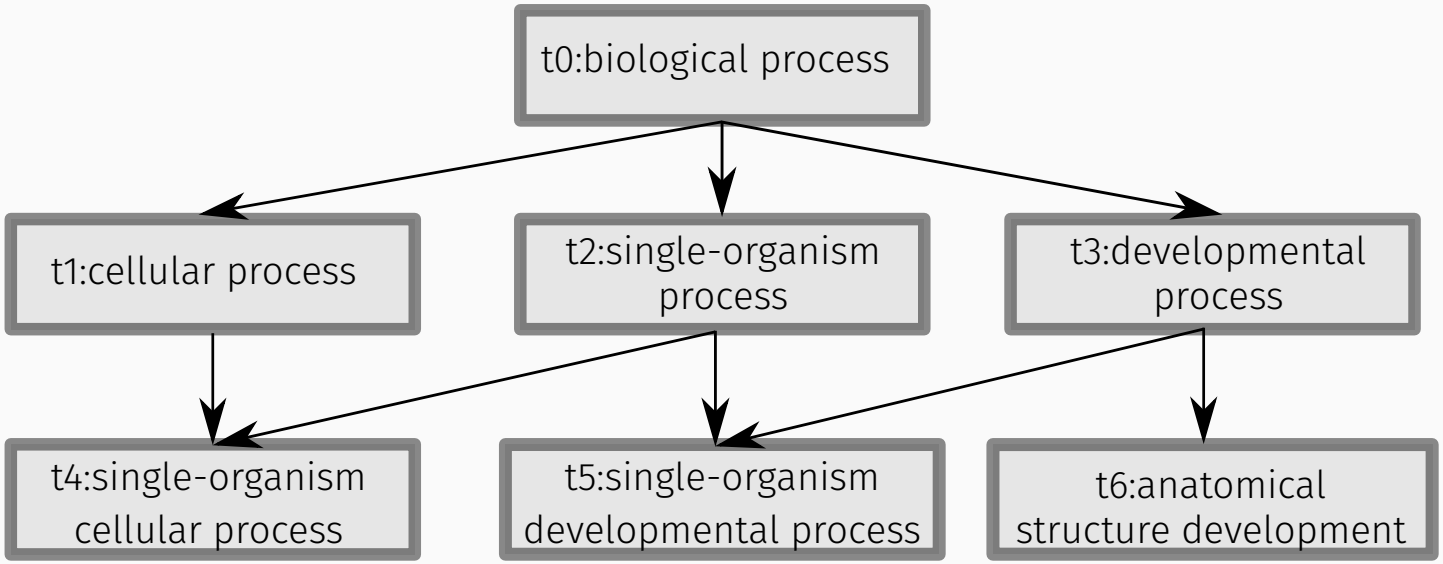
Biclustering (co-clustering, block-clustering)

- Simultaneous partitioning of the set of samples and the set of their features.
- **Outcome** - a set of subsets, i.e. biclusters.



Semantics

- Semantics can be represented by ontologies.
- Ontology as partial-ordered set $< T, \succeq >$, where T represents a set of all terms and $\succeq$ is a binary relation defined on T such that $(g, s) \in \succeq \subseteq T \times T$.
- E.g. *developmental process* is a subtype of *biological process* term, written as (*biological process*, *developmental process*) $\in \succeq$.
- Ontologies are usually very large, GO has more than 40,000 terms.



Semantic biclustering

- **INPUT**: binary gene expression matrix, column and row ontology
- **OUTPUT**: conjunction of column and row terms from corresponding ontology
- **Goal** : detect interpretable rectangular patterns in binary data matrices

```
2 ************************* FINAL RULESET *************************
3 ===== RULE 1=====
4  STATS: score 0.677327 t-score: 90.9458 POSITIVE: 81 NEGATIVE: 324
5  RULE: GO:0005515 AND GO:1901564 AND GO:0048522 AND GO:0005634
6  DETAILS:
7 ID: GO:0005515
8 NAME: protein binding
9 DEF: "Interacting selectively and non-covalently with any protein or protein complex (a complex of two or more proteins that may include other
  nonprotein molecules)." [GOC:go_curators]
10
11 ID: GO:1901564
12 NAME: organonitrogen compound metabolic process
13 DEF: "The chemical reactions and pathways involving organonitrogen compound." [GOC:pr, GOC:TermGenie]
14
15 ID: GO:0048522
16 NAME: positive regulation of cellular process
17 DEF: "Any process that activates or increases the frequency, rate or extent of a cellular process, any of those that are carried out at the
  cellular level, but are not necessarily restricted to a single cell. For example, cell communication occurs among more than one cell, but occurs
  at the cellular level." [GOC:jid]
18
19 ID: GO:0005634
20 NAME: nucleus
21 DEF: "A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of
  the cell's chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialized
  cell types, RNA metabolism or DNA replication may be absent." [GOC:go_curators]
```

Approach 1 - bidirectional enrichment

1. Find a small set of biclusters
   - that cover as many 1s as possible and as few 0s as possible,
   - the bicluster semantics disregarded for the moment,
   - the external PANDA+ tool used for this purpose.
2. Reveal the semantics of biclusters
   - enrichment analysis for finding ontology terms enriched in each bicluster,
   - permutation statistics used to find a proper p-value threshold,
   - topGO R package used for genes and GO terms,
   - Fisher test employed for other ontologies.

Kléma, J.; Malinka, F.; Železný, F. Semantic biclustering for finding local, interpretable and predictive expression patterns. BMC Genomics. 2017

Approach 2 - classification-learning problem

1. It is based on a reduction of the problem to a classification-learning problem.
2. The basic idea is transformation of binary matrix into linearized form
3. Each matrix element represents an example in the form

$$t_1, t_2, \cdots, t_g, t_{g+1}, t_{g+2}, \cdots, t_{g+s}, \textit{expression}$$

   where *expression* indicates expression of gene i for sample j.
4. Classification model is learn to predict expression from $t_1 \cdots, t_{g+s}$ predictors.

Approach 2 - rule and tree learning

- For comparison we used two well-known machine learning methods decision tree (J48) and rule learning (JRip).
- Cluster can be represented as a conjunction of terms

$$\wedge_{k \in G} t_k \wedge_{k \in S} t_{k+g}$$

- JRip — conjunctions of terms $t_1, \cdots, t_{g+s} \rightarrow \textit{expression}$
- J48 — conjunctions of terms as a root-to-leaf paths
- The adjusted conjunctions of these methods are able to represent subsets of matrix the same way as the semantic biclustering.

Results

Method	AUROC	#of biclusters	Avg. # of terms/bicluster
BE	0.823 ± 0.006	11.8 ± 1.5	64.8 ± 3.4
Rules (JRip)	0.636 ± 0.01	102.6 ± 21.5	7.1 ± 0.61
Tree (J48)	0.659 ± 0.01	109.9 ± 5.2	25.4 ± 2.0

Table: Results for Ovary dataset.

Method	AUROC	#of biclusters	Avg. # of terms/bicluster
BE	0.608 ± 0.03	16.4 ± 4.7	47.9 ± 2.13
Rules (JRip)	0.567 ± 0.01	25.9 ± 6.2	7.89 ± 0.53
Tree (J48)	0.627 ± 0.05	20.6 ± 11.09	11.01 ± 4.71

Table: Results for IDICS dataset.

Approach 3 - Rule learning

GOALS:

- speed-up a process of rule induction
- removing redundant terms in rules during the rule induction
- simplify hyperparameter tuning

Outcome:

1. the single rule learning algorithm for the rule induction
2. new feature selection methods
3. new refinement operator

Malinka, F.; Železný, F.; Kléma, J. Finding Semantic Patterns in Omics Data Using Concept Rule Learning with an Ontology-based Refinement Operator. BioData Mining. 2020

Single rule learning & refinement operator

Refinement operator allows to generate new candidates appending a feature/term to the rule.

Ontology-based refinement operator reduces a number of generated refined rules using two novel reduction procedures:

- Redundant Generalization,
- Redundant Non-potential.

Algorithm 1: Single rule learning algorithm

input : O, E^+, E^-, *buildMapping*, *stopCondition*
output: R_{BEST} `// conjunction of selectors`

```
1  R_BEST ← ∅; R_BEST_SCORE ← 0
2  M' ← buildMapping(O, E+, E−)
3  F ← featureConstruction(O)
4  R ← featureSelection(F, E+, M')
5  while R ≠ ∅ or !stopCondition do
6      R_new ← ∅
7      foreach r ∈ R do
8          newCandidates ← refineRule(r, F, O, R_BEST_SCORE, E+ ∪ E−, M')
9          R_new ← R_new ∪ newCandidates
10         (R_BEST, R_BEST_SCORE) ← findBestRule(newCandidates, R_BEST_SCORE, R_BEST)
11     end
12     R ← filterRules(R_new)
13 end
14 return R_BEST
```

Redundant Generalization

Redundant Generalizationt omits candidate rules based on the relation generalization-specialization.

Theorem

If a rule $R1$ contains terms $\{t1, t2\}$ where $t1 \succeq t2$ and $\overline{R1} = \{t2\}$ then

$$\Theta(\overline{R1}) = \Theta(R1)$$

and the rule $R1$ is called a redundant generalization.

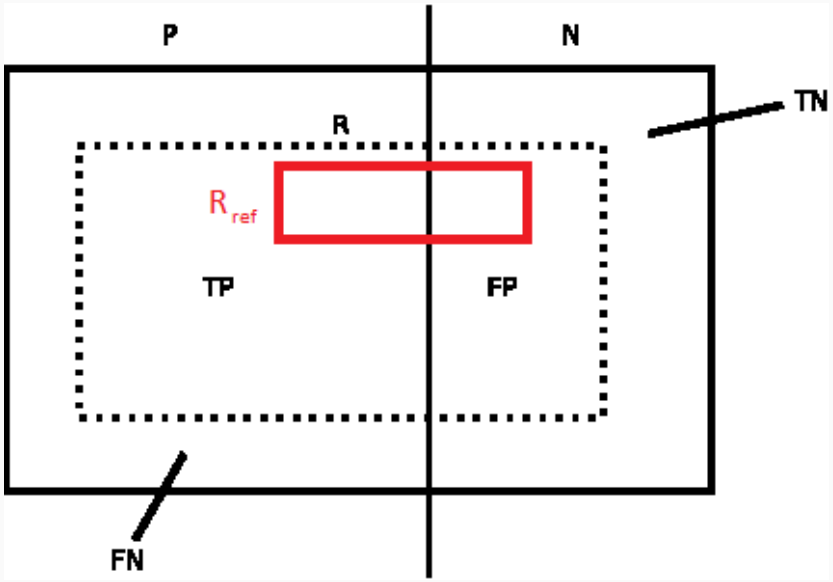
Example

If a rule $R1$ contains terms $\{dog, mammal\}$ where $mammal \succeq dog$ and $\overline{R1} = \{dog\}$ then

$$\Theta(\overline{R1}) = \Theta(R1)$$

Redundant Non-potential

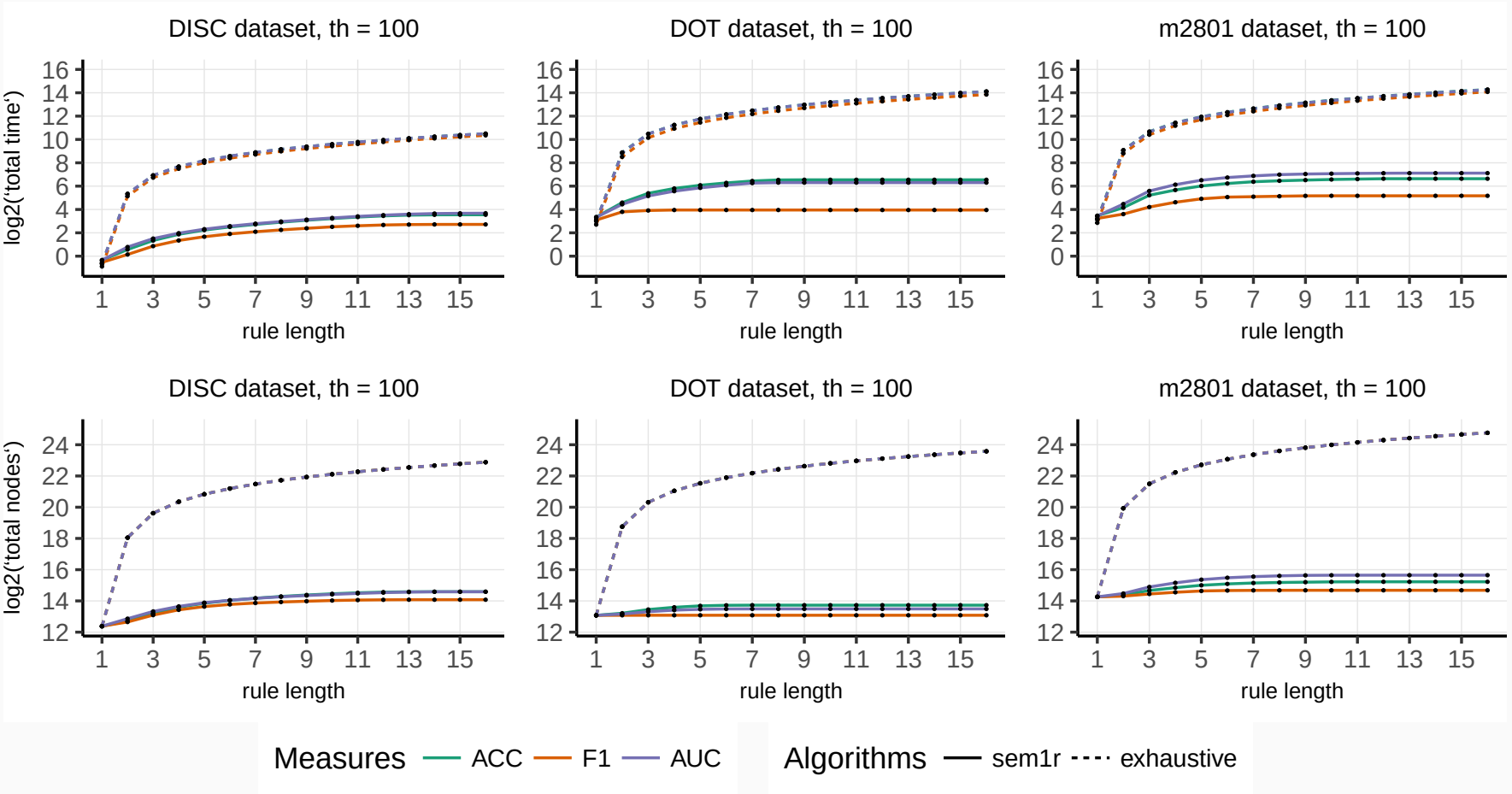
Redundant Non-potential omits the candidate rules which cannot improve classification accuracy.



$ACC(R) = \frac{|TP|+|TN|}{|P|+|N|}$, $ACC_{potential}(R_{ref}) = \frac{|TP|+|TN|+|FP|}{|P|+|N||}$

IF $ACC(R) > ACC_{potential}(R_{ref})$ THEN R_{ref} is Redundant Non-potential

Experiments - graphical results



Conclusions

- The main idea of semantic biclustering is to find biologically interesting and easily interpretable biclusters.
- The concept rule learning algorithm (sem1R) with an ontology-based refinement operator was developed to eliminate particular disadvantages of the previous work.
- The sem1R algorithm in C++ and is available at https://github.com/fmalinka/sem1r as R package.
- We adapted the sem1R algorithm onto two real biological-related problems as:
  1. finding pathogenic variants in a cohort of individuals and
  2. an analysis of E-3 ubiquitin ligase in the gastrointestinal tract.

     Iatsiuk, V., Malinka, F., Pickova, M., Tureckova, J., Klema, J., Spoutil, F., ... Sedlacek, R. (2022). Semantic clustering analysis of E3-ubiquitin ligases in gastrointestinal tract defines genes ontology clusters with tissue expression patterns. BMC gastroenterology