Sémantická dvojshluková analýza dat genové exprese

František Malinka, Jiří Kléma a Filip Železný

Katedra počítačů,
České vysoké učení technické v Praze

IDA
Intelligent Data Analysis
RESEARCH GROUP

ENBIK 2016

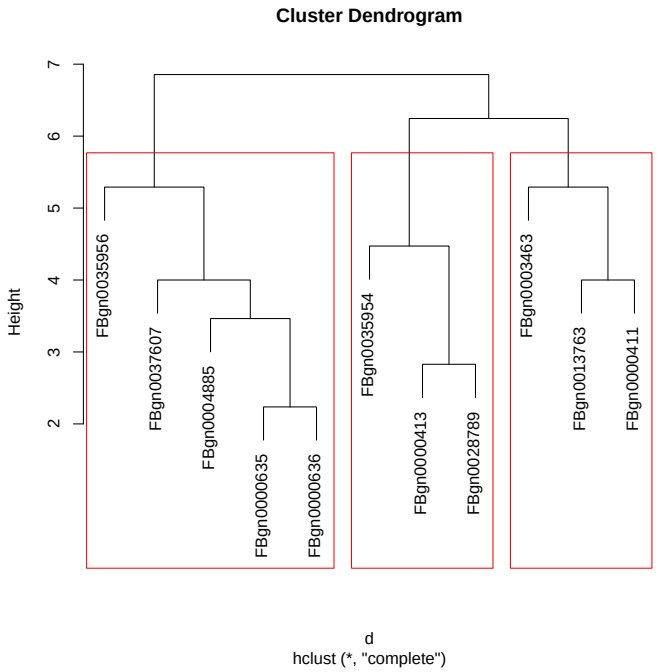
Přehled

- Definování pojmů
  - Shlukování
    - hierarchické shlukování
    - k-means
  - Dvojshlukování
    - motivace, definice, problém složitosti,
    - algoritmy
  - Sémantické dvojshlukování
- Aplikace
  - úloha motivována Dresden ovary dataset,
  - implementace technik symbolického strojového učení

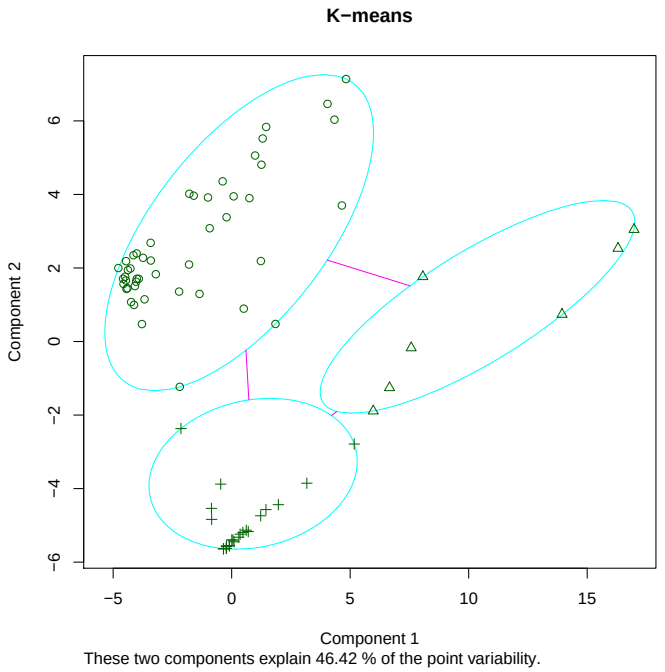
Shlukování (clustering)

- technika pro hledání podskupin (clusterů) v datasetu
- shluk - množina objektů navzájem si podobných a zároveň dostatečně odlišných (nepodobných) od objektů z ostatních shluků
- nutno definovat pojem "podobnost" resp. "odlišnost"

Hierarchické shlukování



K-means



Dvojshlukování (biclustering, co-clustering, block-clustering)

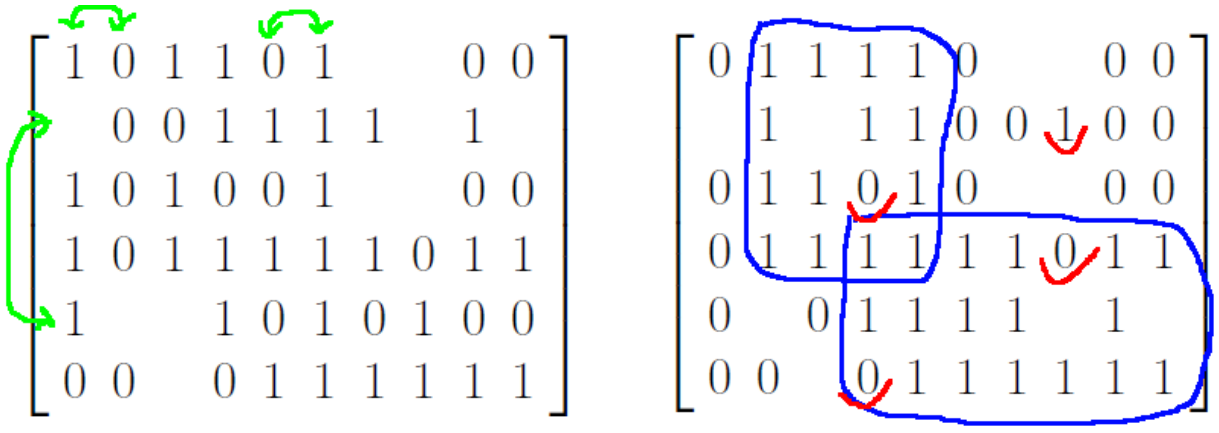
- **souběžné** shlukování přes množinu vzorků a množinu atributů,
- výsledek = identifikace podmnožin vzorků v podmnožině atributů,
- seskupené vzorky a atributy předpokládají vzájemnou relevanci,
- shlukování $\rightarrow$ dvojshlukování $\rightarrow$ multidimenzionální shlukování,

- formálně (z pohledu minimalizace šumu v matici)
  - vstup: $\mathbb{A}^{m \times n}$, $a_{ij} \in \{0, 1\}$ (obecně může být matice reálných čísel),
  - výstup: $\Pi^* = \{P_1, \ldots, P_{|\Pi|}\}$, $P_i = (P_i^t, P_i^f)$, $P_i^t \in \{0,1\}^m$, $P_i^f \in \{0,1\}^n$
  - taková, že: $\Pi^* = \arg\min_\Pi ||\mathcal{N}|| = \arg\min_\Pi \bigvee_{P_i \in \Pi} \left(P_i^t \otimes P_i^f \right) \oplus \mathbb{A}$

  - $\mathcal{N} \in \{0,1\}^{m \times n}$ ... matice šumu (FP a FN),
  - $\oplus$ ... element-wise XOR, $\vee$ ... element-wise OR, $\otimes$ ... outer product,

Dvojshlukování jako hledání přibližných binárních vzorů

- objevování malých množin čtvercových vzorů, které nejlépe reprezentují vstupní matici $\mathbb{A}$,
  - nejlepší reprezentace = kompaktní a deskriptivní
  - greedy algoritmus, začíná se vzorem s nízkým šumem, pak expanduj
  - přidej jako vzor, pokud zvyšuje pokrytí matice $\mathbb{A}$,
  - řízeno tzv. cost function, např. $\rho \sum_{P \in \Pi}(||P^t|| + ||P^f||) + ||\mathcal{N}||$
- PANDA$^+$ [Lucchese et al.,2014] řeší v $\mathcal{O}(rkm^2n)$.

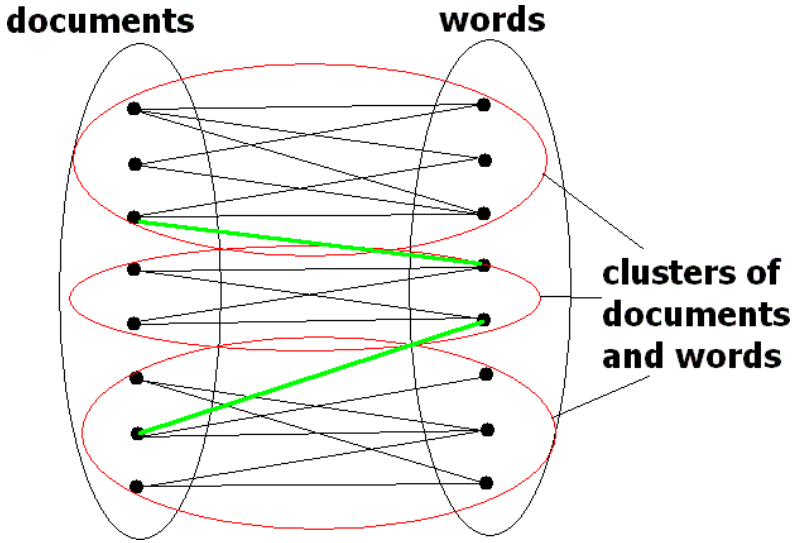


a) původní matice

b) dvojshluky získané minimalizací šumu

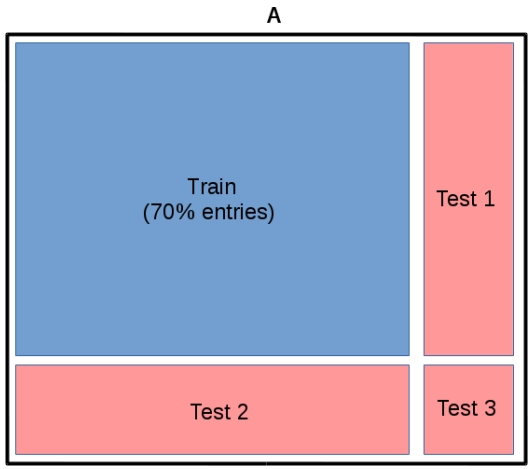
Další metody řešení dvojshlukování

- bipartitní graf a jeho rozklad,
  - NP-úplná, lze užít heuristik
- kombinace výsledků shlukování (*Interralated Two-way Clustering*),
- spektrální dvojshlukování,
- bayesova inference (*Bayesian Biclustering model*),
- a mnoho dalších ...



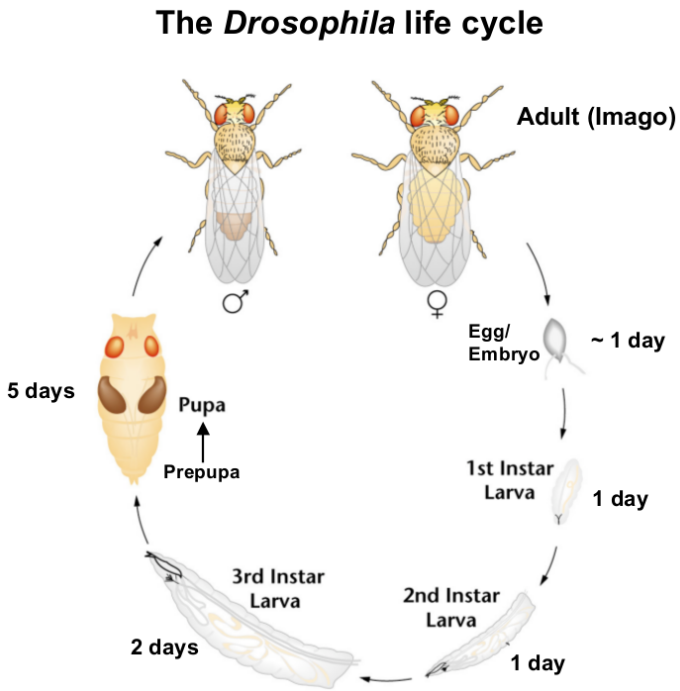
Sémantické dvojshlukování

- sémantické shlukování
  - konvenční shlukování + popis shluků termy s využitím apriorní znalosti,
  - interpretace shluků,
- sémantické dvojshlukování
  - nový koncenpt, sémantické shlukování + dvojshlukování,
  - hlavní evaluační kritérium – přesnost predikce na testovacích datech.



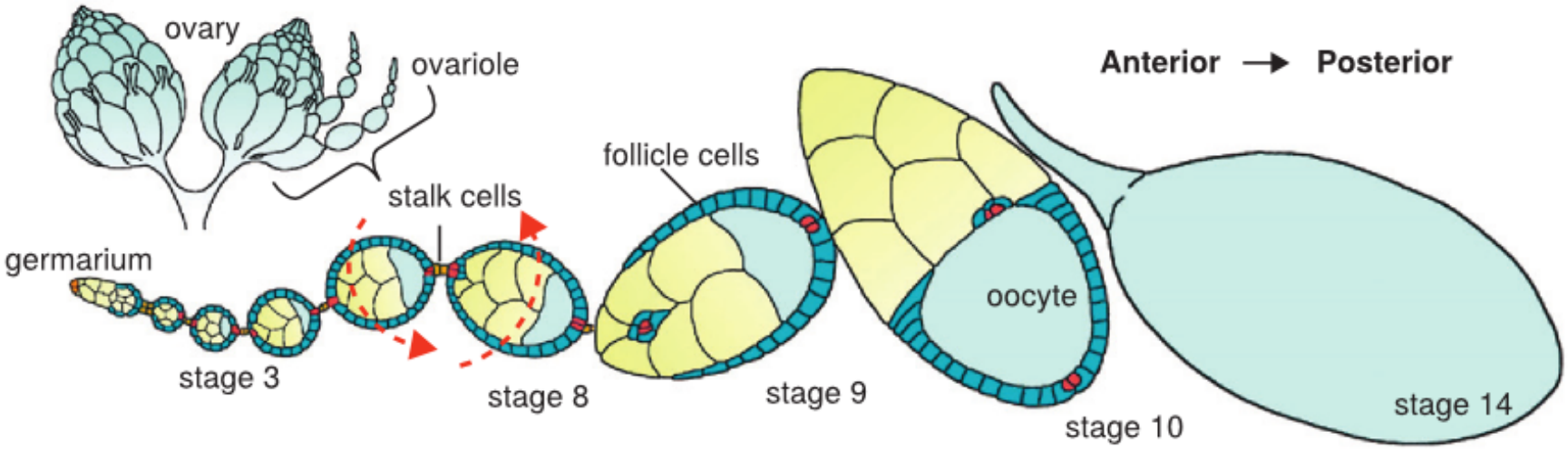
Drosophila melanogaster

- 3mm dlouhá, krátký životní cyklus (zhruba 2 týdny).
- nenároční chov, mnoho potomků,
- v porovnání s lidmi
  - sdílí mnoho konzervovaných genů,
  - podobnosti v základní buněčné struktuře a funkci, ...



Oogeneze

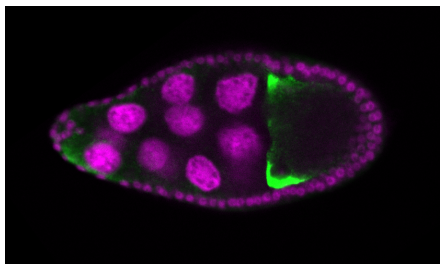
- oogeneze - proces tvorby vajíčka (ovum)
- pár vaječníků (ovarium), každý z nich se skládá ze zhruba 15-18 ovariol
- celkem 14 vývojových fází, buňky jsou kontinuálně diferencovány



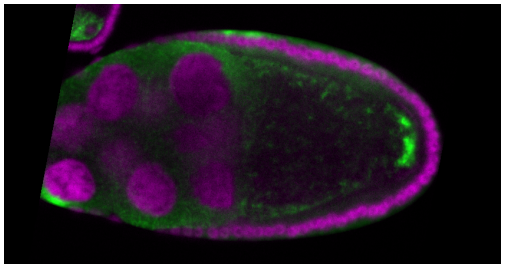
Anatomie vaječníku Drosophily melenogaster - lineární sekvence vývojových fází.

Dresden Ovary Table

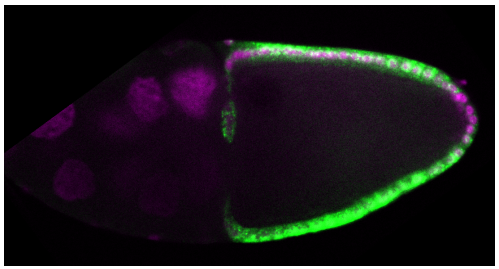
- Dataset stage/location specifických mRNA distribucí
- 2 dimenze, 2 ontologie
  - 6510 genů, Gene ontology
  - 131 stage/location specifických termů, Location ontology
- Formát dat: matice $\mathbb{A}$
  - dimenze $m \times n$, kde řádky jsou geny a sloupce jsou lokace
  - Gene ontology G, Location ontology L
  - element $a_{i,j} \in \{0, 1\}$ indikuje (1) expresi genu i v lokaci j



a) oocyte anterior restriction



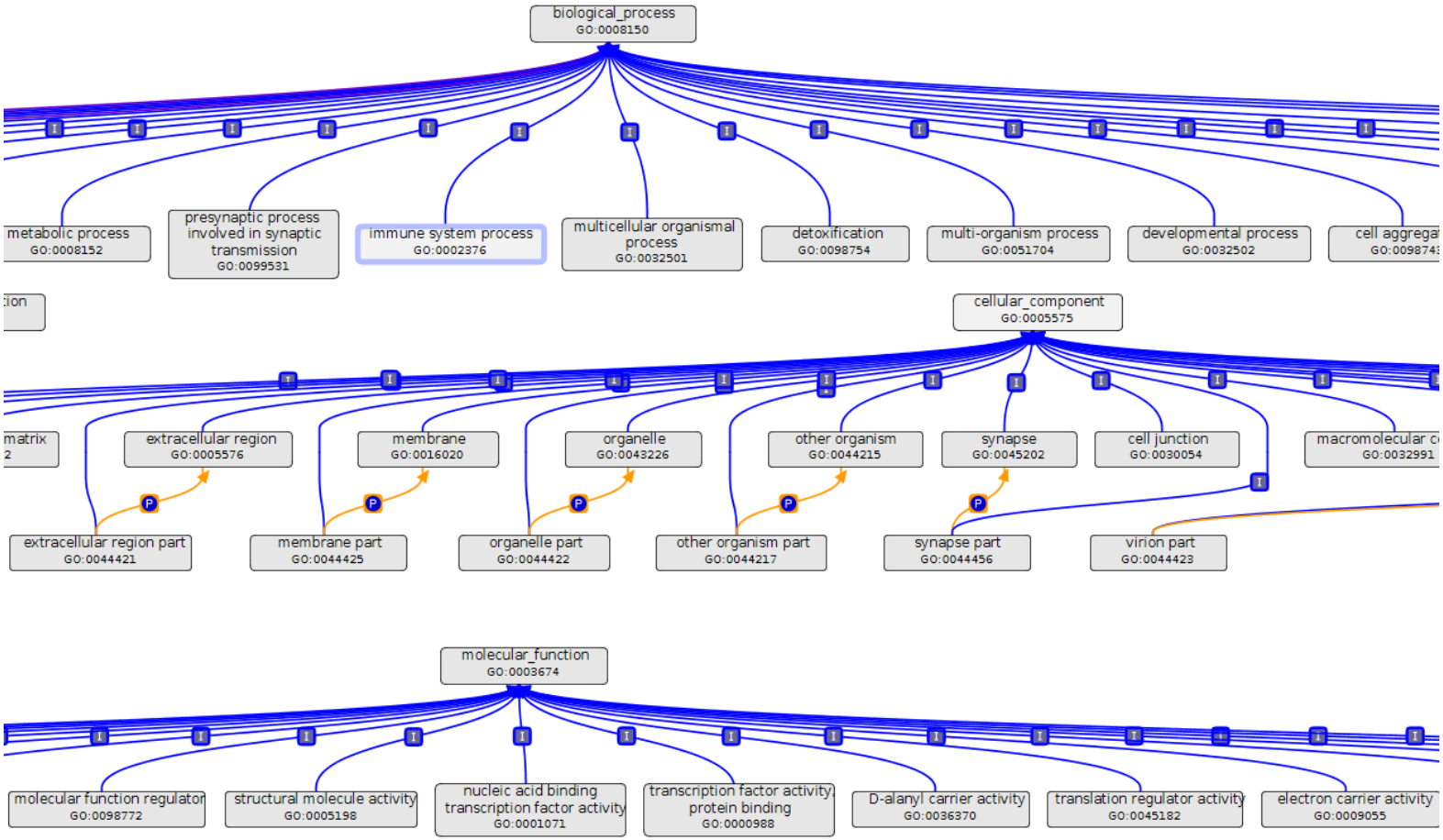
b) oocyte posterior restriction



c) follicle cells

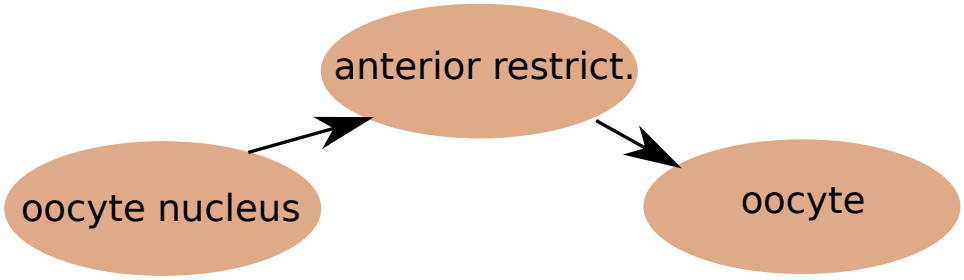
Gene Ontology

- Ontologie definuje koncepty/třídy popisující funkci genů a vztahy mezi nimi.
- Popis genů pomocí: *molecular function*,*cellular component*,*biological process*.

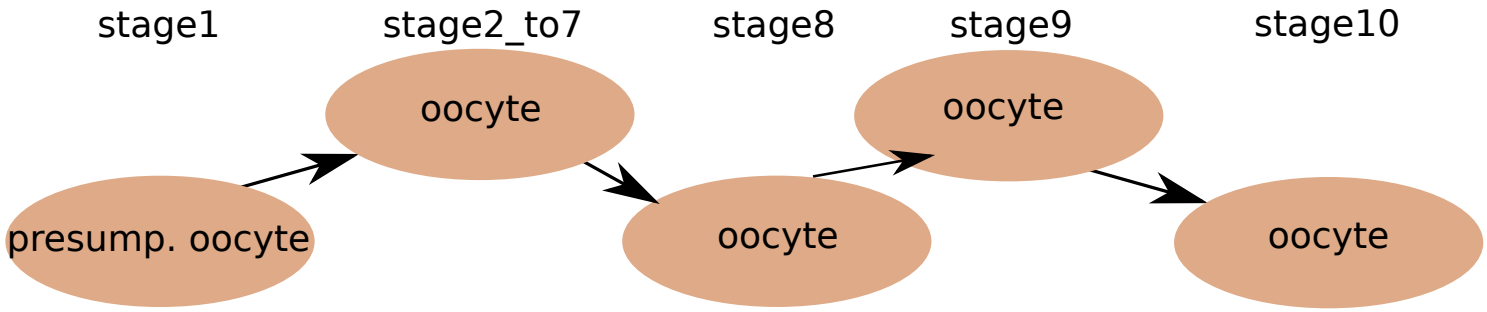


Location Ontology

- hierarchicky kontrolovaný slovník s více jak 100 termy,
- současně popisuje prostorové (*part of*) a časové (*develops from*) vlastnosti.
- *part of*



- *develops from*



Maticová reprezentace

- VSTUP
  - matice genové exprese $\mathbb{A}^{m \times n}$, $a_{i,j} \in \{0, 1\}$, gene ontology G a location ontology L
- VÝSTUP
  - konjunkce lokačních a GO termů z korespondující ontologie
- CÍL
  - najít submatici obsahující zajímavé vzory

Sémantické dvojshlukování

- jako úloha symbolického strojového učení založená na redukci problému na klasifikační úlohu
- Základní myšlenkou je transformace matice $\mathbb{A}$ na její linearizovanou formu.
- Každý element matice $a_{i,j}$ reprezentuje jeden příklad ve formě

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

  kde $expression$ indikuje expresi genu i pro lokaci j.
- Klasifikační model je učen k predikci $expression$ z $t_1 \ldots, t_{g+s}$ prediktorů.

Rule a tree learning metoda

- Použity dobře zavedené metody strojového učení *decision tree* (J48) and *rule learning* (JRip).
- Dvojshluk může být reprezentován jak konjunkce termů

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

- JRip — konjunkce termů $t_1, \ldots, t_{g+s} \rightarrow expression$
- J48 — konjunkce termů jako cesta od kořene k listu

Transformace na trénovací dataset

Algorithm 2: Unrolling $\mathbb{A}$ into $\mathbb{B}$.

input : $\mathbb{A}^{m \times n}$, $a_{i,j} \in \{0, 1\}$
output: $\mathbb{B}^{m \cdot n \times g+s+1}$, $b_{i,j} \in \{0, 1\}$

```
1  /* Genes are represented by a set of FBgn identifiers                */
2  M ← getAllGeneNames(A); // all genes in A
3  G ← getAllGoTerms(M); // GO transitive closure wrt M
4  g ← |G|;
5  for i ← 1 to m do
6      ∀x ∈ {1, ..., g + s + 1} : T_x ← 0; // initialization
7      for j ← 1 to g do
8          if term g_j is associated with gene m_i then
9              T_j ← 1
10         end
11     end
12     for k ← 1 to s do // where s is a set of situation terms
13         associations ← find all associations in a set of situations for s_k;
14         for ∀assoc ∈ associations do
15             T_{g+assoc_i} ← 1; // where assoc_i is an index of situation term
               assoc
16         end
17         T_{g+s+1} ← a_{i,k}; // add expression indicator
18         B_{i,*} ← T;
19     end
20 end
21 B ← filterGoTerms(B, Θ)// due to a given threshold Θ;
```

Výsledky

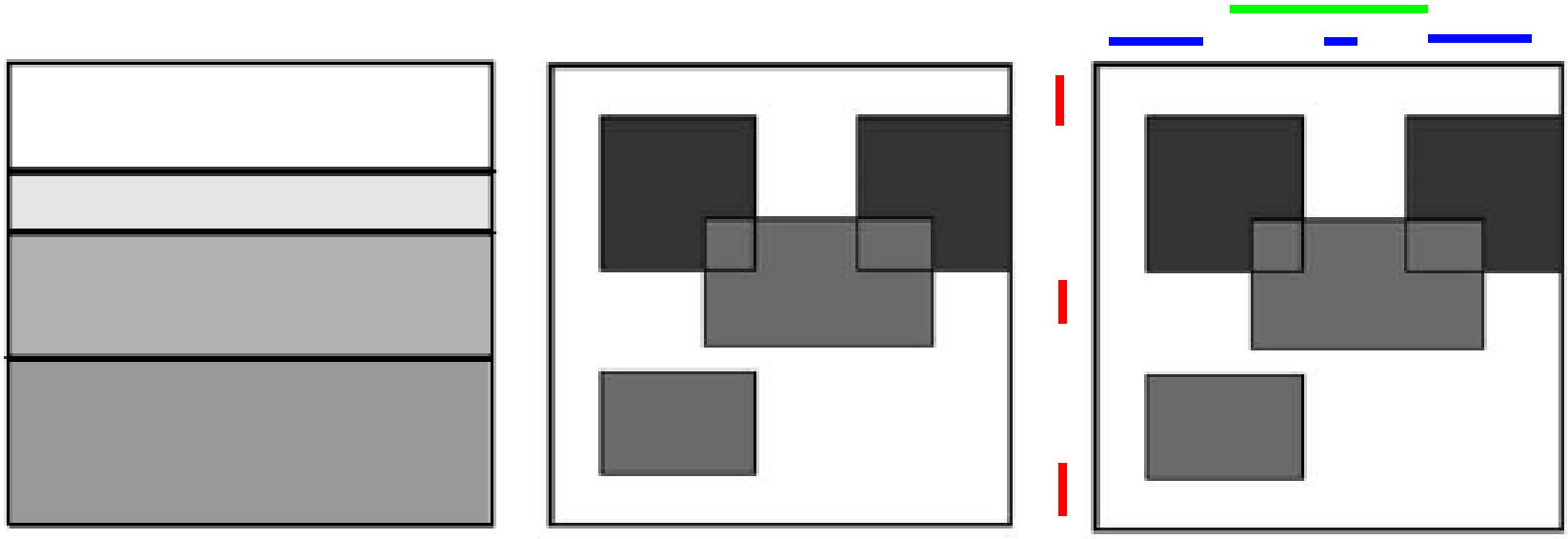
Method	AUROC	#of biclusters	Avg. # of terms/bicluster
Bicluster Enrichment	0.769 ± 0.013	11.8 ± 1.5	47.9 ± 2.13
Rules (JRip)	0.636 ± 0.01	93.7 ± 17.4	7.0 ± 0.40
Tree (J48)	0.713 ± 0.01	1 ± 0	27.5 ± 0.89

Table 1: Výsledky pro dataset s Gene a location ontologií.

Method	AUROC	#of biclusters	Avg. # of terms/bicluster
Rules (JRip)	0.567 ± 0.01	25.6 ± 7.28	7.69 ± 0.52
Tree (J48)	0.723 ± 0.01	1 ± 0	23.66 ± 1.21

Table 2: Výsledky pro DISK dataset s Gene a DAO ontologií.

Shrnutí



a) clustering

b) biclustering

c) semantic biclustering