Laboratory of Information Systems
Institute of Applied Mathematics and
Computer Science at Tula State University

Mikhail Bogatyrev
okkambo@mail.ru

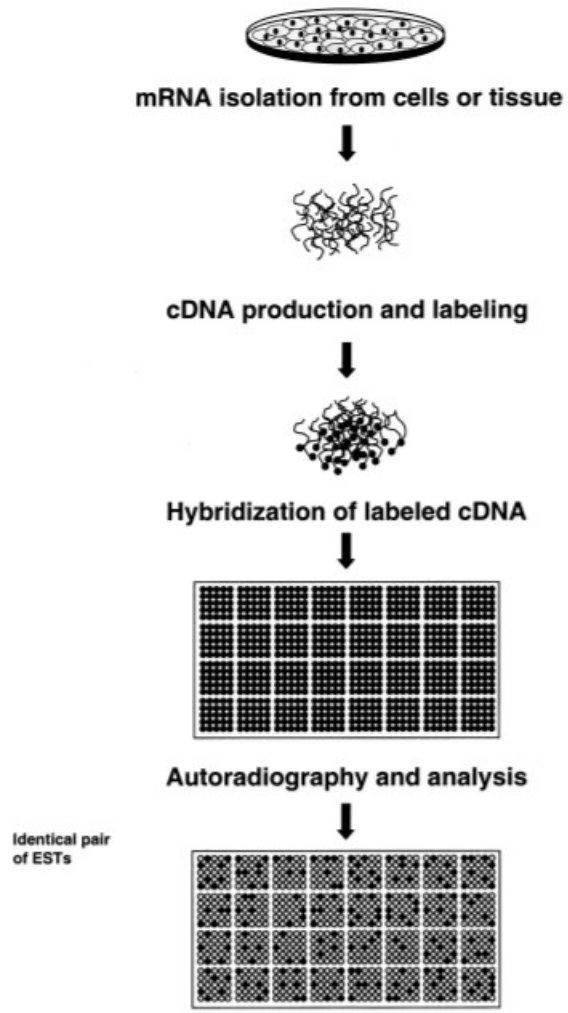
Clustering Methods in Gene Expression Research

ICMBB-20
Pushchino

Outline

- The task of gene expression analysis
- Gene expression data (GED) organization
- Conventional GED clustering
- Bi- and triclustering methods in the GED clustering
- Evolutionary approach to GED clustering
- Conceptual clustering with Formal Concept Analysis

Gene expression analysis



The study of the way genes are transcribed to synthesize functional gene products — functional RNA species or protein products.

Elements of Microarray Technology.

Chip manufacture: A microarray is a small chip onto which tens of thousands of DNA molecules are attached in fixed grids. Each grid cell relates to a DNA sequence.

Target preparation, labeling and hybridization: Typically, two mRNA samples (a test sample and a control sample) are reverse-transcribed into cDNA (targets), labeled using either fluorescent dyes or radioactive isotopics, and then hybridized with the probes on the surface of the chip.

The scanning process: Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets.

The problem of clustering

Clustering is the division of a set of objects into disjoint subsets – clusters, so that in each of the subsets the objects are as close as possible to each other according to a certain criterion, and their intercluster proximity is minimal.

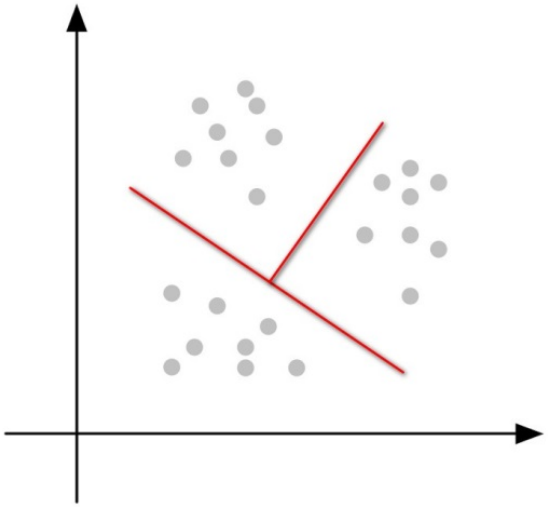


Fig.1

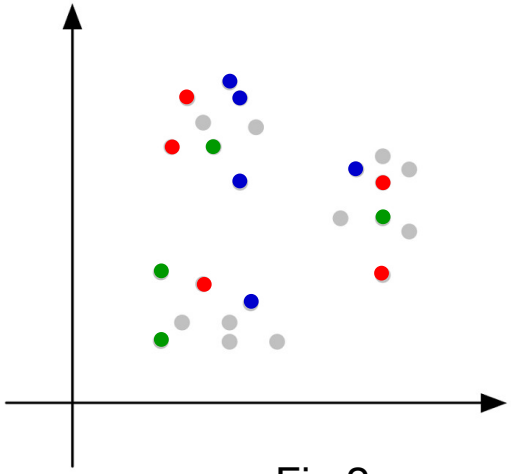
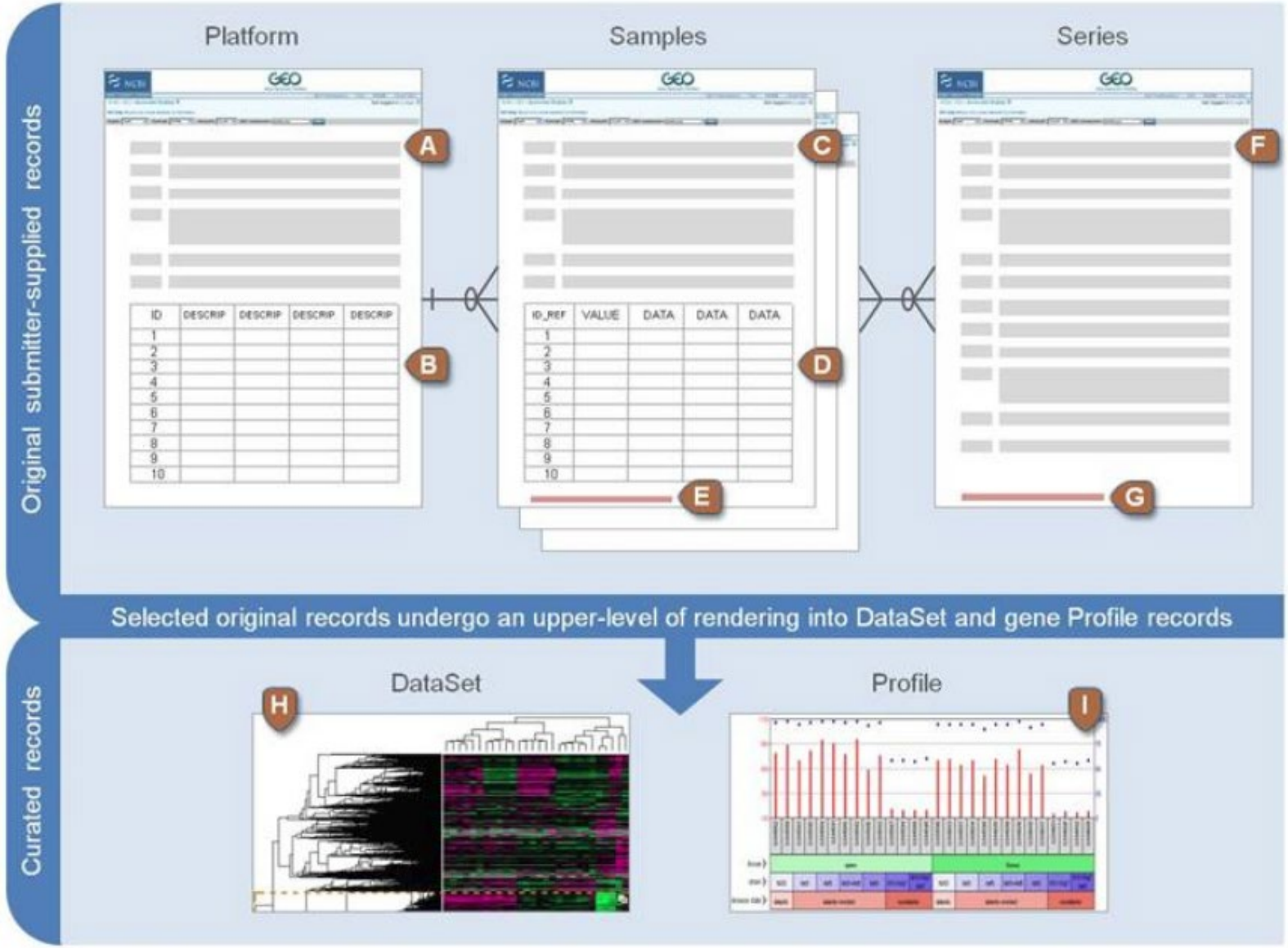


Fig.2

Relativity of the proximity measure. Points on the Fig.1 can be separated into 3 clusters according with the Euclidean proximity measure based on calculating of point coordinates. On the Fig. 2 the proximity measure is color and "colored" clusters intersect according with the Euclidean proximity measure.

Gene expression data organization



Example of Gene Expression Data (GED)

Data on *p* genes for *n* samples:

mRNA samples

M	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Genes (Spots)

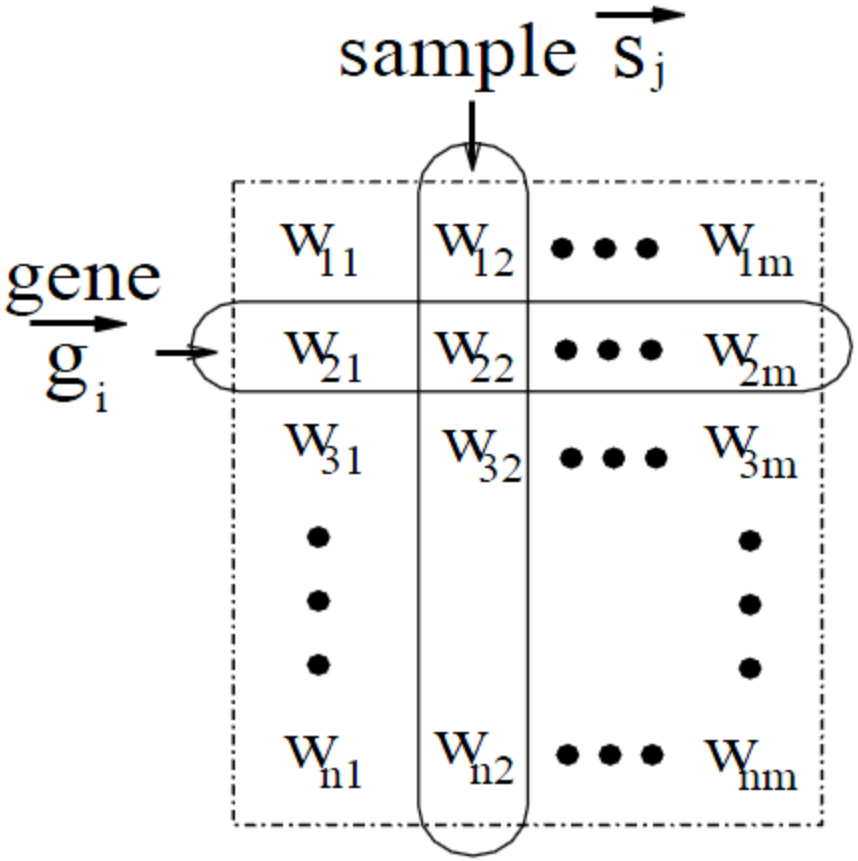
Gene expression level of gene *i* in mRNA sample *j*

= (normalized) Log_2(Red intensity / Green intensity)

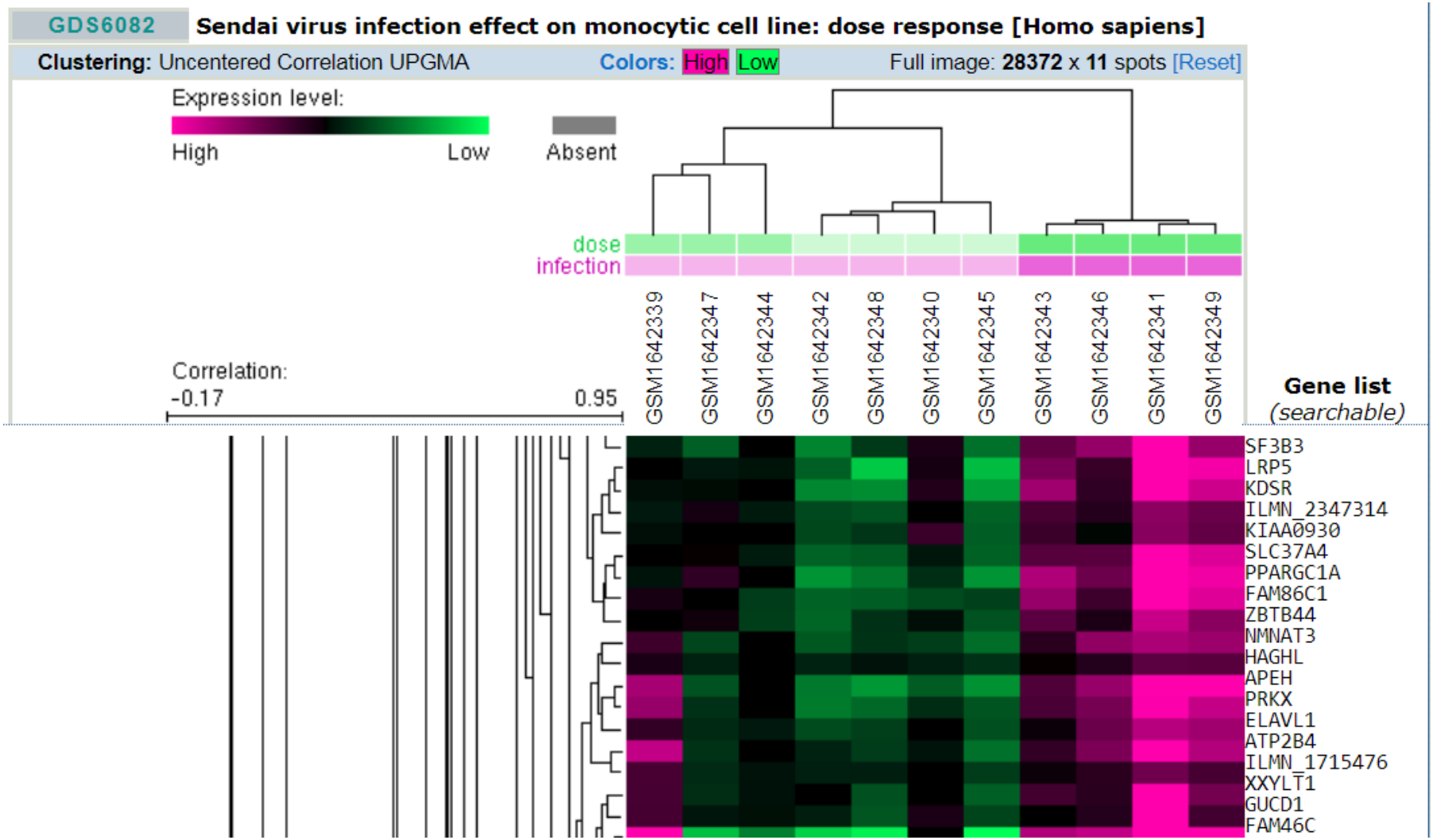
Conventional clustering of GED

sample-based clustering regards the samples as the objects and the genes as the features.

gene-based clustering: the genes are treated as the objects, while the samples are the features



GED clustering in GEO



Traits of GED conventional clustering

Problem of proximity measurement.

- Euclidean distance $E(\vec{g}_i, \vec{g}_j) = \sum_{k=1}^{m} (\vec{g}_{ik} - \vec{g}_{jk})^2$

- Pearson's correlation coefficient

$$P(\vec{g}_i, \vec{g}_j) = \frac{\sum_{k=1}^{m} (\vec{g}_{ik} - \mu_{g_i})(\vec{g}_{jk} - \mu_{g_j})}{\sqrt{(\vec{g}_{ik} - \mu_{g_i})^2}\sqrt{(\vec{g}_{jk} - \mu_{g_j})^2}}$$

Traits of GED conventional clustering

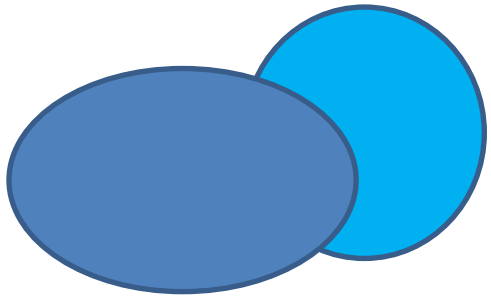
Problem of the choice of clustering algorithm

Classical
- *k*-means algorithm
- **Hierarchical clustering algorithms**

Non-classical
- Self organizing maps
- *Evolutionary clustering algorithms*

Resume:

Clusters overlap



Reasons for biclustering

1) Grouping of genes according to their expression under multiple conditions.

2) Classification of a new gene, given its expression and the expression of other genes, with known classification.

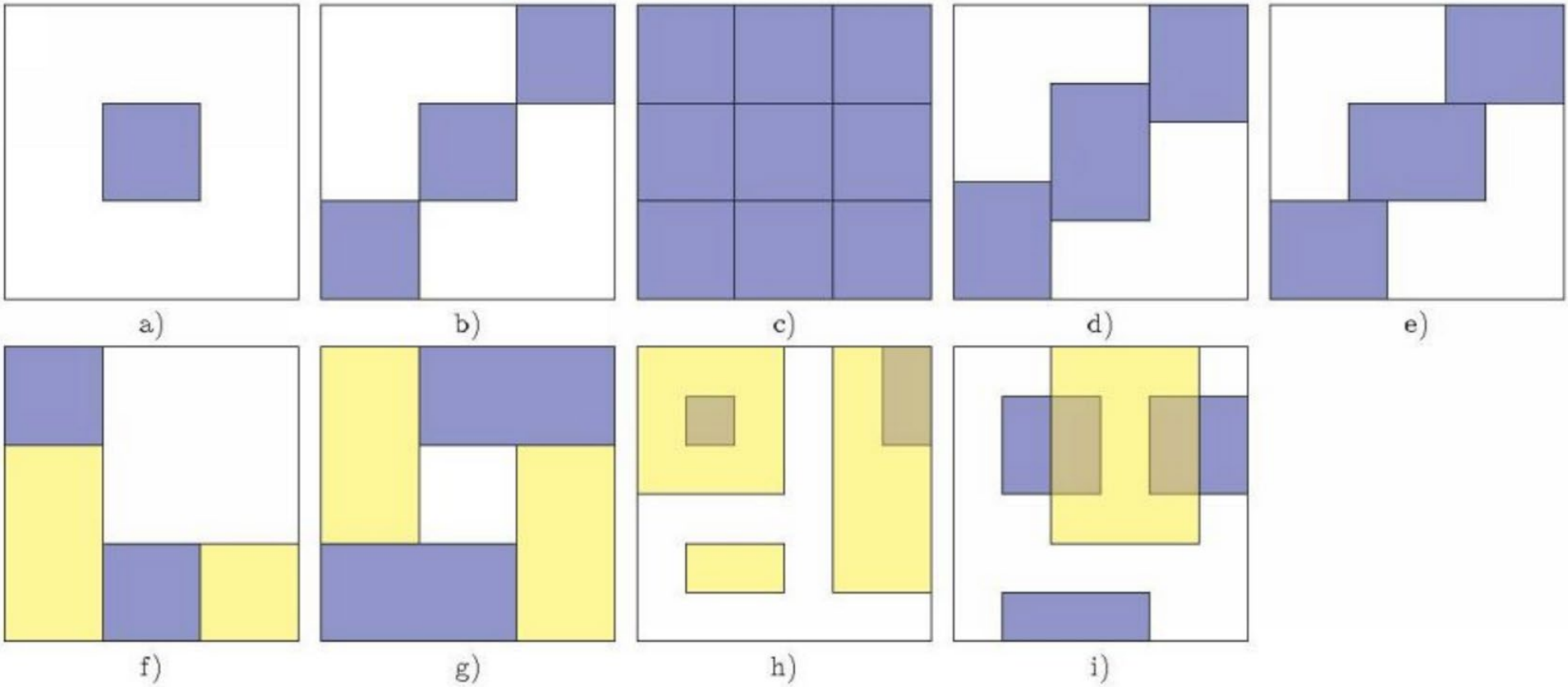
3) Grouping of conditions based on the expression of a number of genes.

4) Classification of a new sample, given the expression of the genes under that experimental condition.

"Gene expression data is meaningful when clustering both genes and samples".

Biclustering

Examples of biclusters



Biclusters of GED

S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1

(f) Overall Coherent Evolution

S1	S1	S1	S1
S2	S2	S2	S2
S3	S3	S3	S3
S4	S4	S4	S4

(g) Coherent Evolution on the Rows

S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4

(h) Coherent Evolution on the Columns

70	13	19	10
29	40	49	35
40	20	27	15
90	15	20	12

(i) Coherent Evolution on the Columns

Biclustering techniques

Problem Complexity: the problem known to be NP-complete

- graph-theoretic approach

  Weighted Bipartite Graphs

- statistical approach

S.C. Madeira ; A.L. Oliveira. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics Vol.: 1 , Issue: 1 , Pp. 24 - 45

Formal Concept Analysis

1. Formal Context:

$K = (G, M, R)$ $\quad$ G – set of objects

M – set of attributes

$R \subseteq G \times M$

R – relation

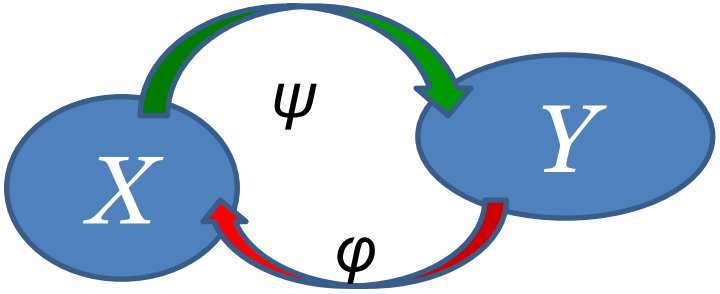
Example

	Membrane	Nucleus	Replication	Recombination
DNA				X
Virus				X
Prokaryotes	X		X	
Eukaryotes	X	X	X	
Bacterium	X		X	

FCA as biclustering technique

2. Formal Concept:

$$(X, Y) \quad X \subseteq G,\ Y \subseteq M$$



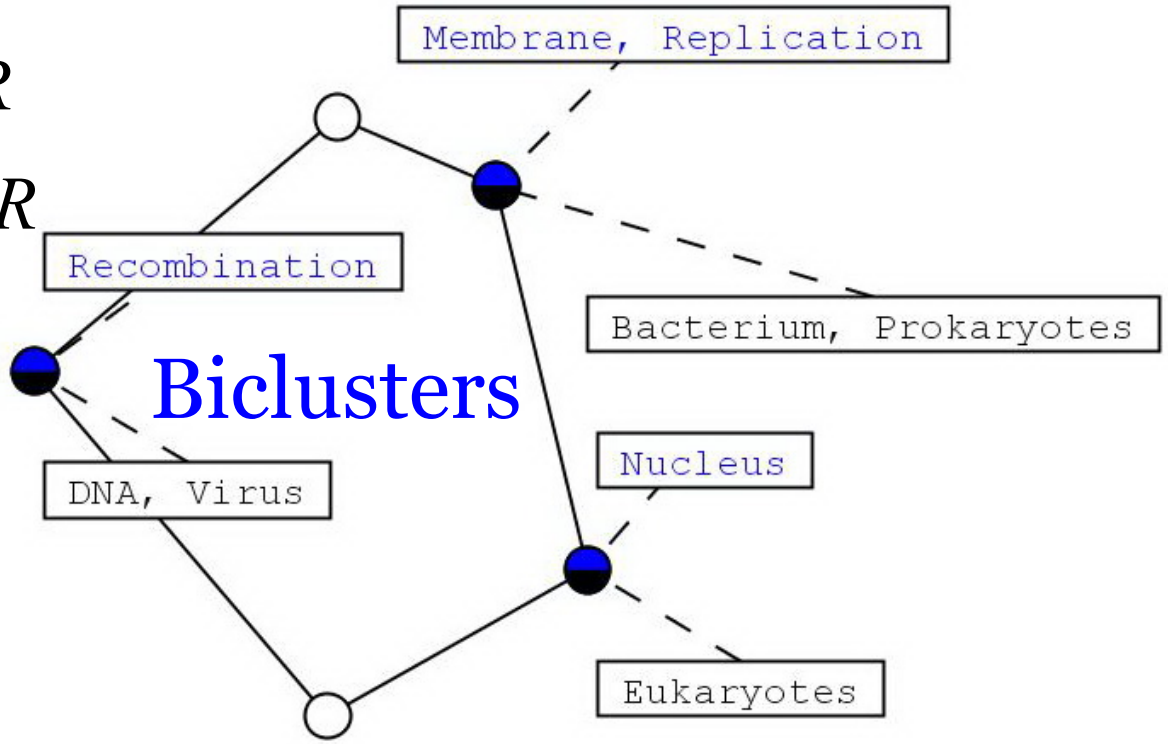
Galois Connection (Mapping)

$$X \xrightarrow{\psi} Y \qquad Y \xrightarrow{\varphi} X$$

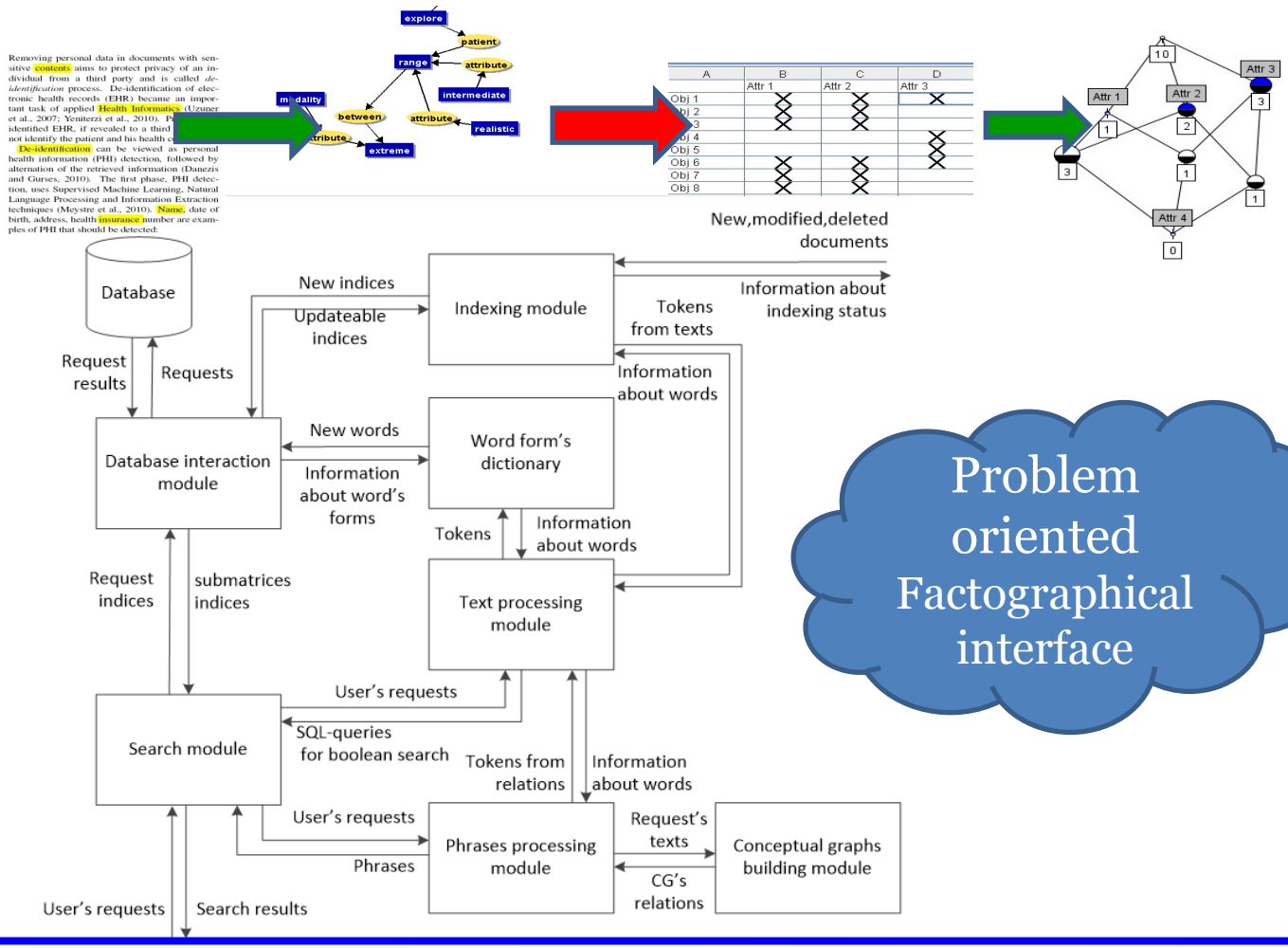
$$\forall x \in O < x, \psi(x) > \in R$$

$$\forall y \in A < \phi(y), y > \in R$$

3. Conceptual Lattice:



FCA Fact Extraction Environment



Queries

Facts

Facts inside clusters

Problem oriented Factographical interface

Mikhail Bogatyrev, Kirill Samodurov. Framework for Conceptual Modeling on Natural Language Texts. – Proc. of the International Workshop on Concept Discovery in Unstructured Data (CDUD 2016). CEUR workshop proceedings. Vol-1625. Pp. 13-24. 2016

Biclustering GED with FCA

Methods:

- conceptual scaling [1]
- pattern structures [2]
- interval method of K-FCA [3]

1) Khalid Raza. FormalConceptAnalysisforKnowledge Discovery fromBiological Data. t: https://www.researchgate.net/publication/277603473.
2)M. Kaytoue, S.O. Kuznetsov, A. Napoli, S. Duplessis. Mining gene expression data with pattern structures in formal concept analysis. Information Sciences, V. 181, Issue 10, 2011. Pp. 1989-2001.
3) JoseMGonzález-Calabozo,FranciscoJValverde-AlbaceteandCarmenPeláez-Moreno. Interactiveknowledgediscoveryanddata miningongenomicexpressiondatawith numericformalconceptanalysis BMCBioinformatics (2016) 17:374

Triclusters & tricioncepts in FCA

$\mathrm{K} = (G, M, B, R)$ Triadic formal context

$R \subseteq G \times M \times B$ R - relation

$T = (X, Y, Z) = (g^{\square}, m^{\square}, b^{\square})$

$X \subseteq G, Y \subseteq M, Z \subseteq B$ Tricluster

$$\rho(X, Y, Z) = \frac{|R \cap X \times Y \times Z|}{|X||Y||Z|}$$ Density of tricluster

If $\rho(X, Y, Z) = 1$ then tricluster is triconcept

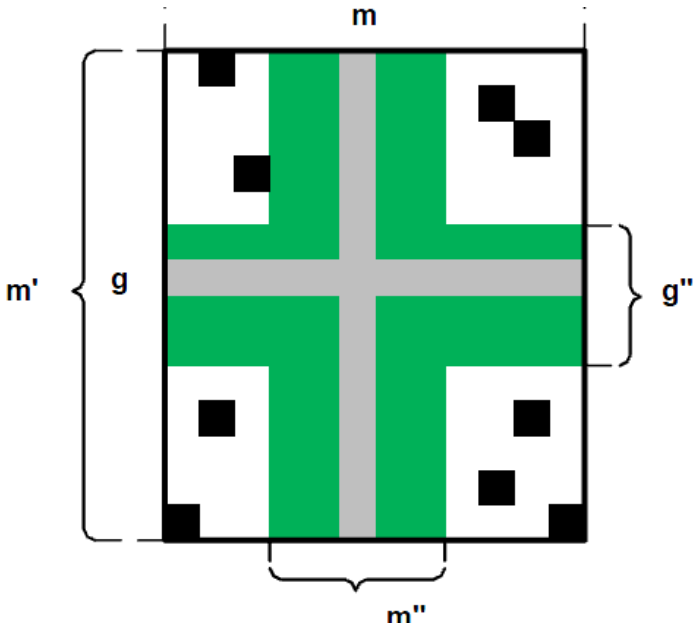
OAC-triclustering algorithm`s operators

$m' = \{ (g, b) \mid (g, m, b) \in Y \}$	$m'' = \{ \tilde{m} \mid (g, b) \in m' \quad \text{and} \quad (g, \tilde{m}, b) \in Y \}$
$g' = \{ (m, b) \mid (g, m, b) \in Y \}$	$g'' = \{ \tilde{g} \mid (m, b) \in g' \quad \text{and} \quad (\tilde{g}, m, b) \in Y \}$
$b' = \{ (g, m) \mid (g, m, b) \in Y \}$	$b'' = \{ \tilde{b} \mid (g, m) \in b' \quad \text{and} \quad (g, m, \tilde{b}) \in Y \}$

$$g^{\square} = \{ g_i \mid (g_i, b_i) \in m' \text{ or } (g_i, m_i) \in b' \}$$

$$m^{\square} = \{ m_i \mid (m_i, b_i) \in g' \text{ or } (g_i, m_i) \in b' \}$$

$$b^{\square} = \{ b_i \mid (g_i, b_i) \in m' \text{ or } (m_i, b_i) \in g' \}$$



FCA-OLAP triclustering

OLAP: On-Line Analytical Processing

OLAP operators

SLICE()
DICE()
PIVOT()
DRILL()

SAP Sybase IQ

Intellectual Queries

S. Ferre, P. Allard, O. Ridoux. Cubes of Concepts: Multi-dimensional Exploration of Multi-valued Contexts. – Proc. ICFCA-2012. LNAI, v. 7278. Pp. 112-127. Springer-Verlag, 2012

Resume

Analysis of clustering methods in the study of gene expression leads to the following conclusions.

1. There are several directions of clustering application, which use the features of expression data and set specific tasks of its study. These are conventional and multimodal clustering, evolutionary clustering and clustering using Formal Concept Analysis.
2. The trend towards unification of data storage and software solutions for their processing leads to the emergence of integrated data warehouses such as GEO, which open up new opportunities in the study of gene expression. Such capabilities include, for example, the use of machine learning on GED with the next formation of a forecast of their state. Modern neural network technologies allow solving similar problems on big expression data.

Acknowledgments. The reported study was funded by Russian Foundation of Basic Research according to research project № 19-07-01178 and RFBR and Tula Region according to research project № 19-47-710007.

Thank you for your attention.