Integrative analysis of large-scale biological data sets

Enrico Glaab, Jonathan M. Garibaldi, Natalio Krasnogor January 2011

Outline

Overview:

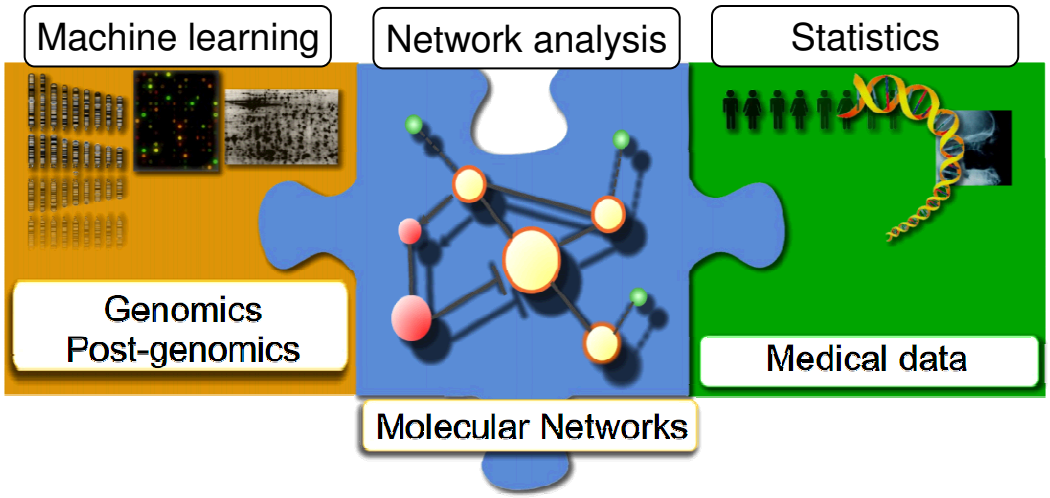
- **Introduction**: main goals and data sets
- **ArrayMining.net**: tool set for integrative microarray analysis
- **TopoGSA**: network topological analysis of genes/proteins
- **Conclusions**

Introduction

Research questions and goals behind the thesis

- **Typical problem in biosciences**: How to make effective use of multiple, large-scale data sources?
- **Typical problem in computer science**: How to exploit the strengths of different algorithms for the same/related purpose?

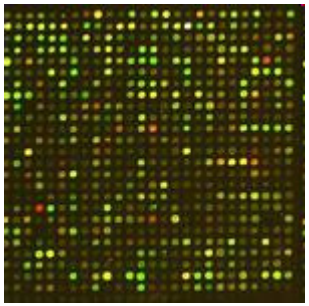
→ **GOAL**: Develop new methods combining diverse data sources and algorithms



Introduction (2): Our data

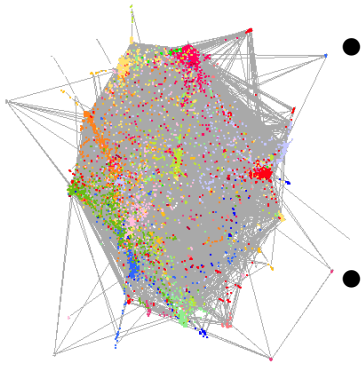
Biological data sources used:

Breast cancer microarray data:



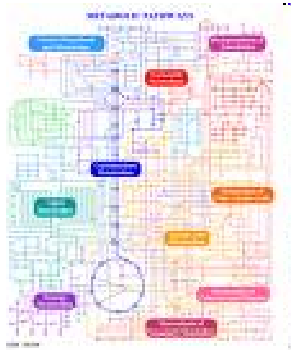
- public microarray data sets: Huang et al., Veer et al.
- pre-processing: GC-RMA

Protein interaction data:



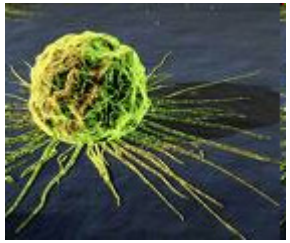
- unweighted binary interactions (MIPS, DIP, BIND, HPRD, IntAct - only human)
- 9392 nodes, 38857 edges

Cellular pathway data:



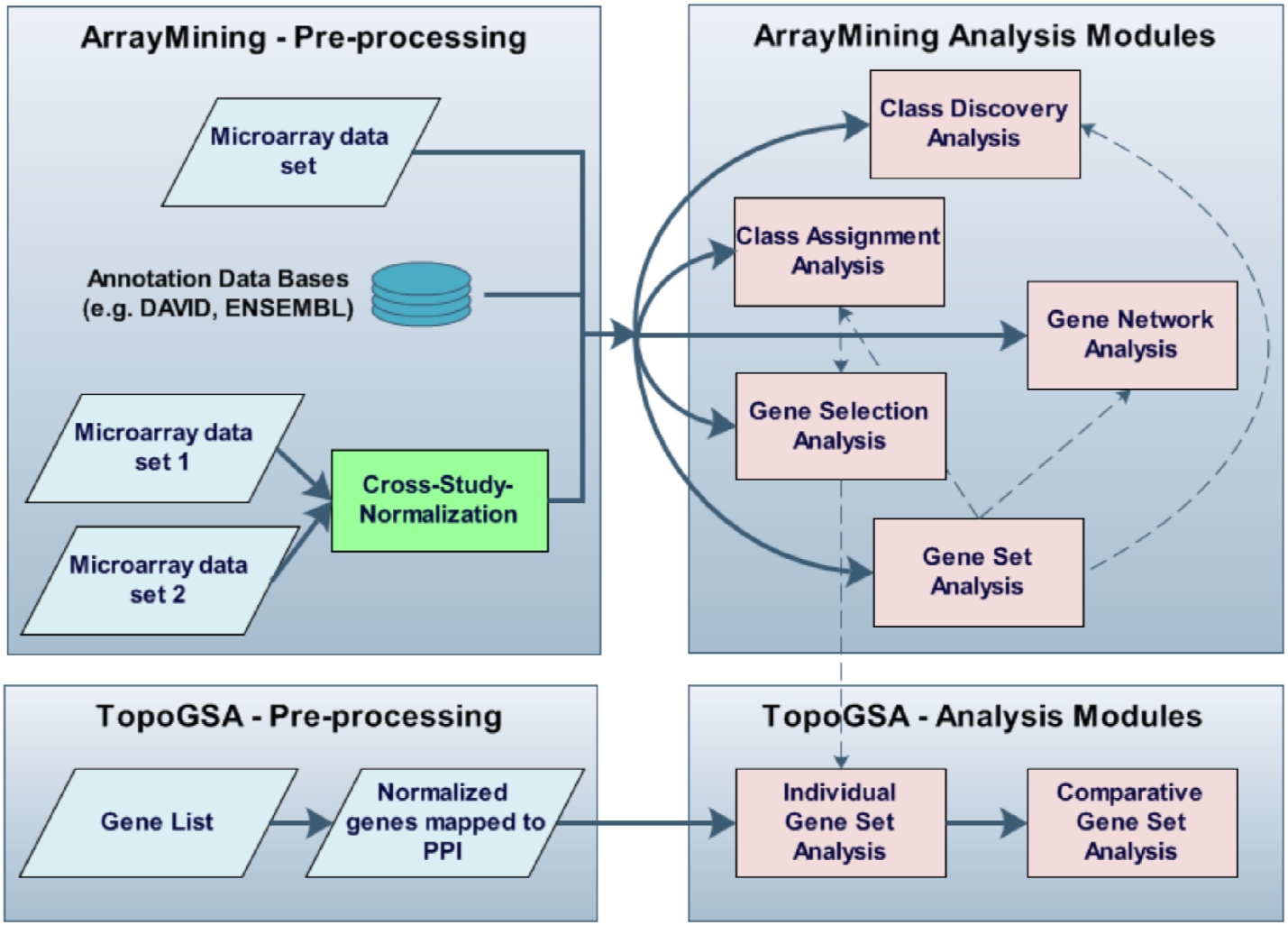
- obtained from GO, BioCarta, Reactome, KEGG and InterPro
- total: approx. 3000 pathways (size > 10)

Cancer gene sets:



- mutated genes in different human cancer types (Breast, Liver,...)
- 30 gene sets of size > 10 genes

Methods overview: ArrayMining & TopoGSA



Web-tool: ArrayMining.net

www.arraymining.net

Goal: A "swiss knife" for microarray analysis tasks

What is ArrayMining.net?

ArrayMinining.net is an online microarray analysis tool set integrating multiple data sources and algorithms.

6 analysis modules:

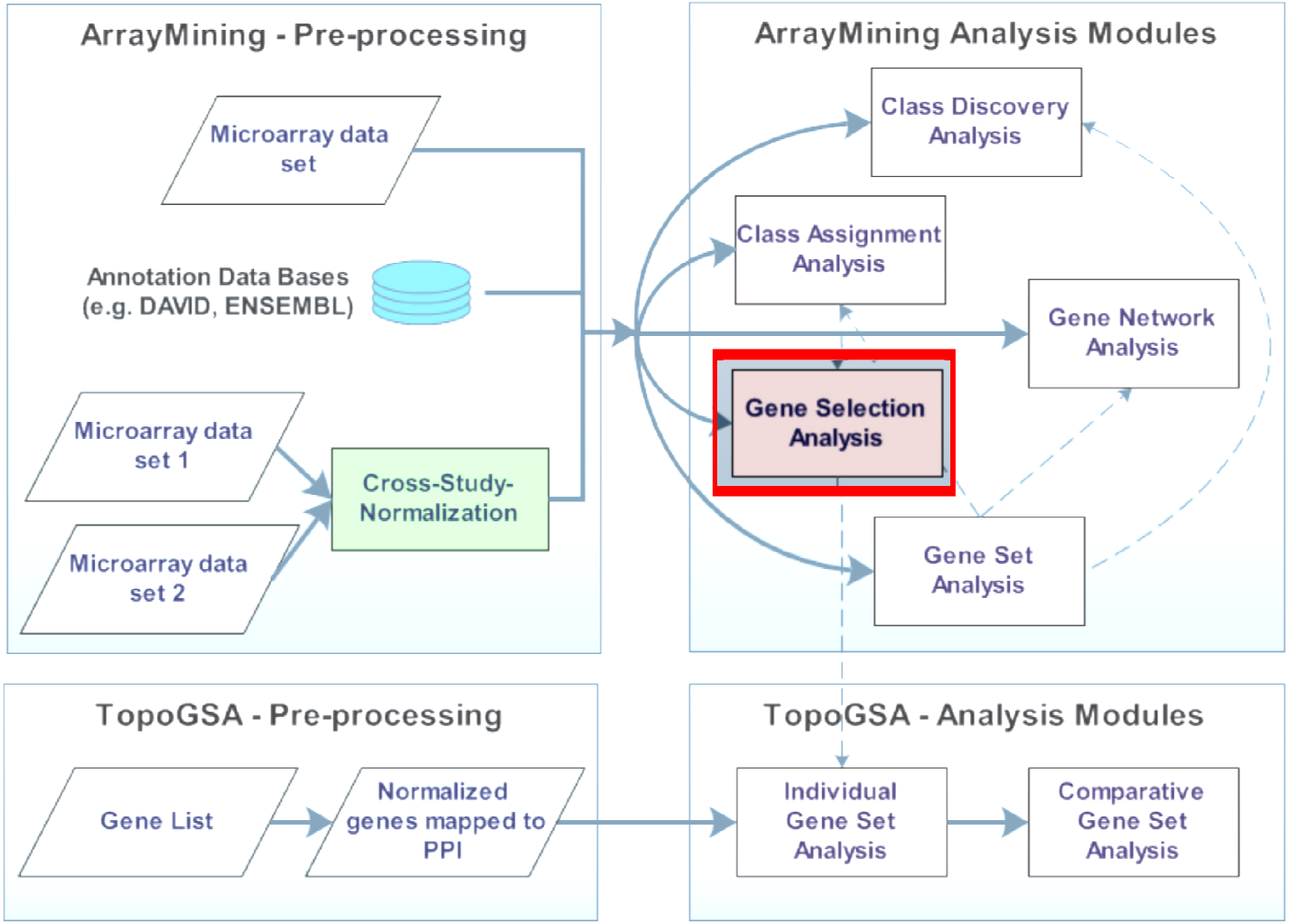
1. Gene selection
2. Sample clustering
3. Sample classification

} classical

4. Gene Set Analysis
5. Gene Network Analysis
6. Cross-Study Normalization

} new

Methods overview: ArrayMining & TopoGSA

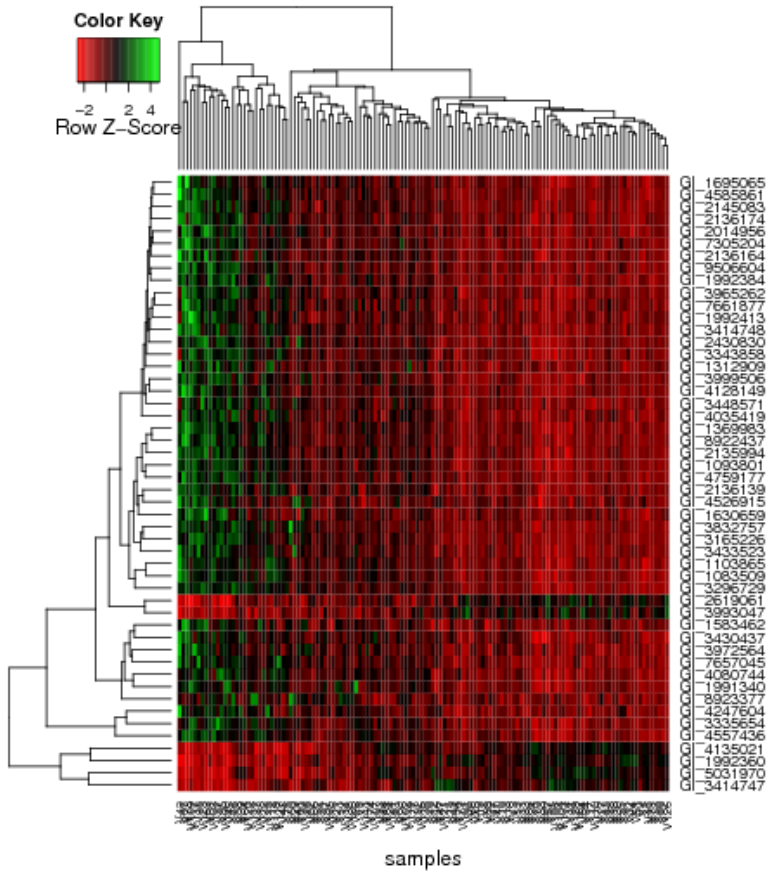


Gene selection: QMC Breast cancer data set

- all top-ranked genes are known or likely to be involved in breast cancer
- the selection is robust with regard to cross-validation cycles and algorithms

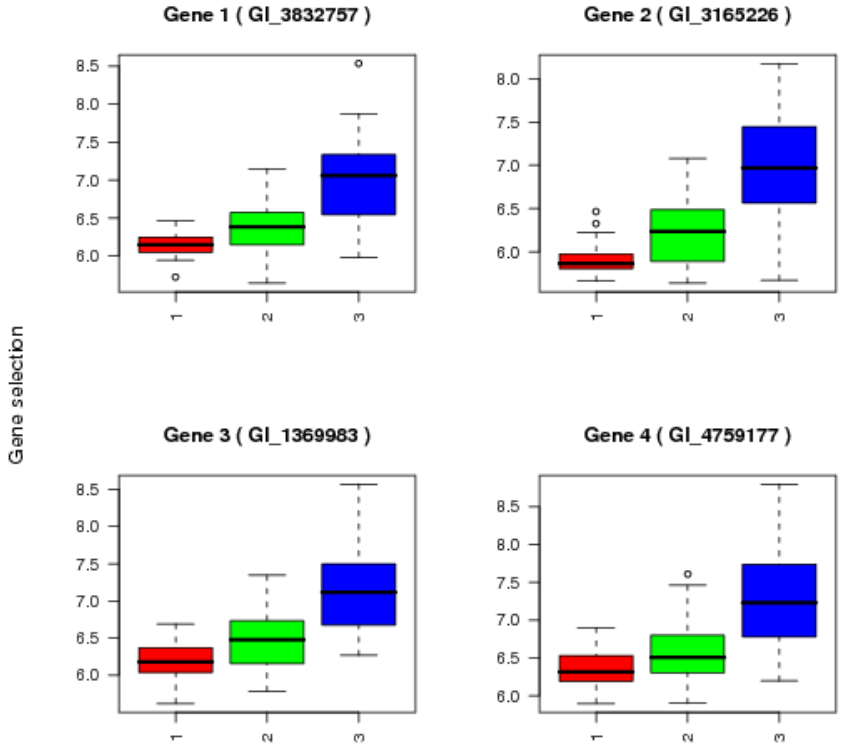
Gene name	PC (gene vs. outcome):	Fold Change	Q-value (Rank)
ESTROGEN RECEPTOR 1	-0.75	0.16	1.6e-20 (1.)
RAS-LIKE, ESTROGEN-REGULATED, GROWTH INHIBITOR	-0.66	0.46	5.3e-14 (2.)
WD REPEAT DOMAIN 19	-0.66	0.73	1.2e-13 (3.)
CARBONIC ANHYDRASE XII	-0.65	0.28	2.7e-13 (4.)
ARP3 ACTIN-RELATED PROTEIN 3 HOMOLOG (YEAST)	0.64	1.37	9.6e-13 (5.)
TETRATRICOPEPTIDE REPEAT DOMAIN 8	-0.63	0.82	2.2e-12 (6.)
BREAST CANCER MEMBRANE PROTEIN 11	-0.62	0.24	7.1e-12 (7.)

Visualization of results: QMC Breast cancer data



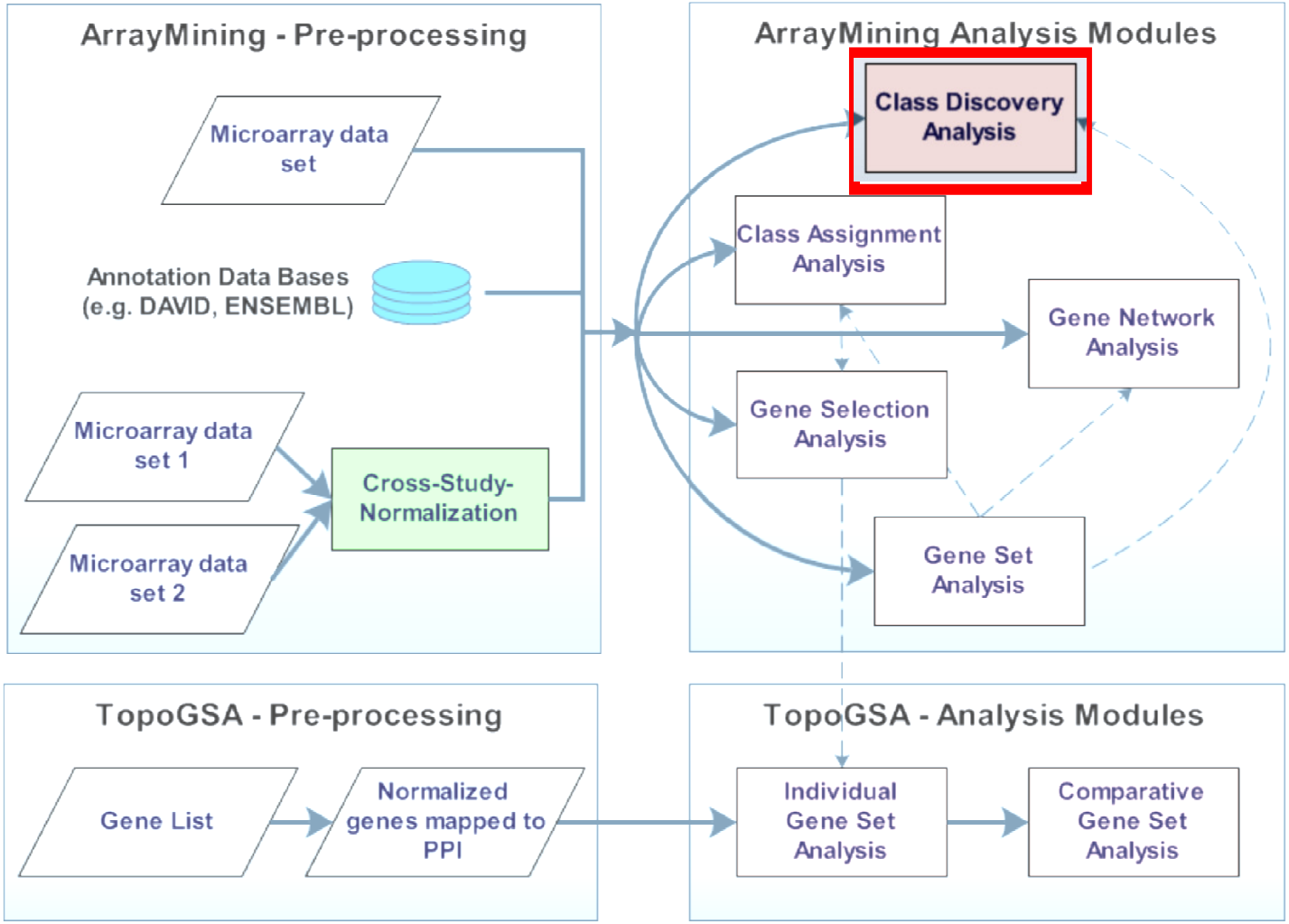
Heat map: 50 most significant genes

Expression levels across 3 tumour grades:



Box plot: 4 most significant genes

Methods overview: ArrayMining & TopoGSA



ArrayMining.net: Example

ArrayMining - Class Discovery Analysis module:

- **Motiviation**:
  Exploiting the synergies between partition-based and hierarchical clustering algorithms

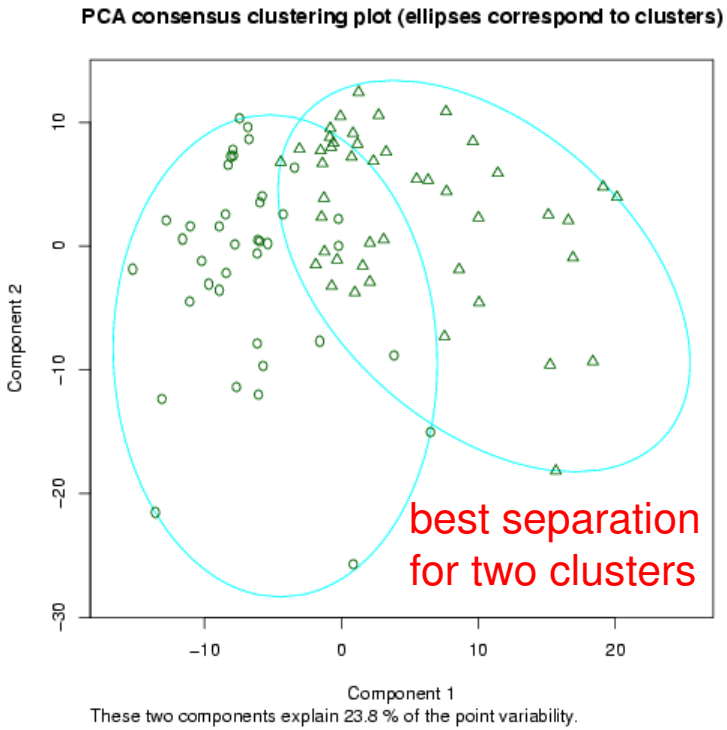
- **Approach:**
  Consensus clustering based on the agreement of clustering results for pairs of objects (details on next slide).
  - equivalent to median partition problem (NP-complete)
  - Simulated Annealing (SA) has been shown to provide good solutions

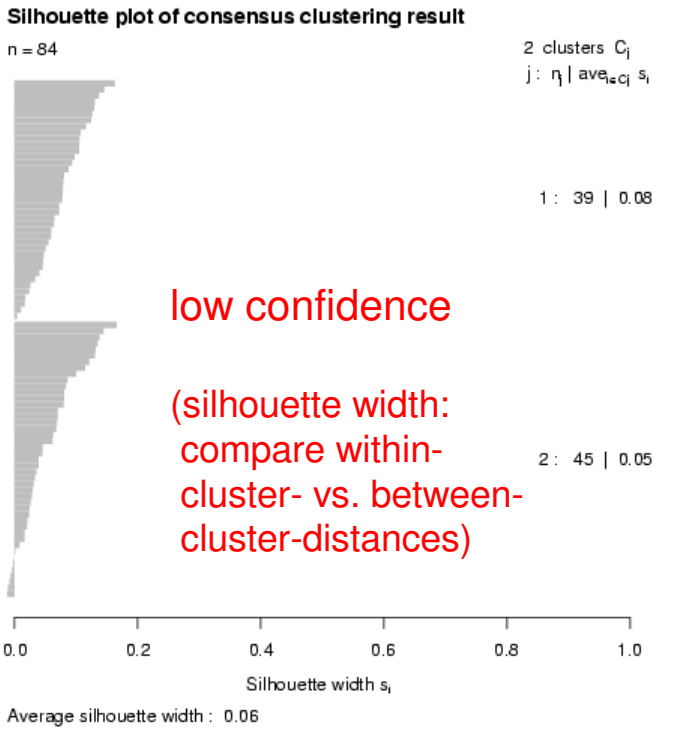
- **Our solution**:
  - Compare SA (Aarts et al. cooling scheme) with thermodynamic SA (TSA) and fast SA (FSA) → FSA provides fastest convergence
  - Initialization: Input clustering with highest agreement to other inputs

Consensus clustering: example

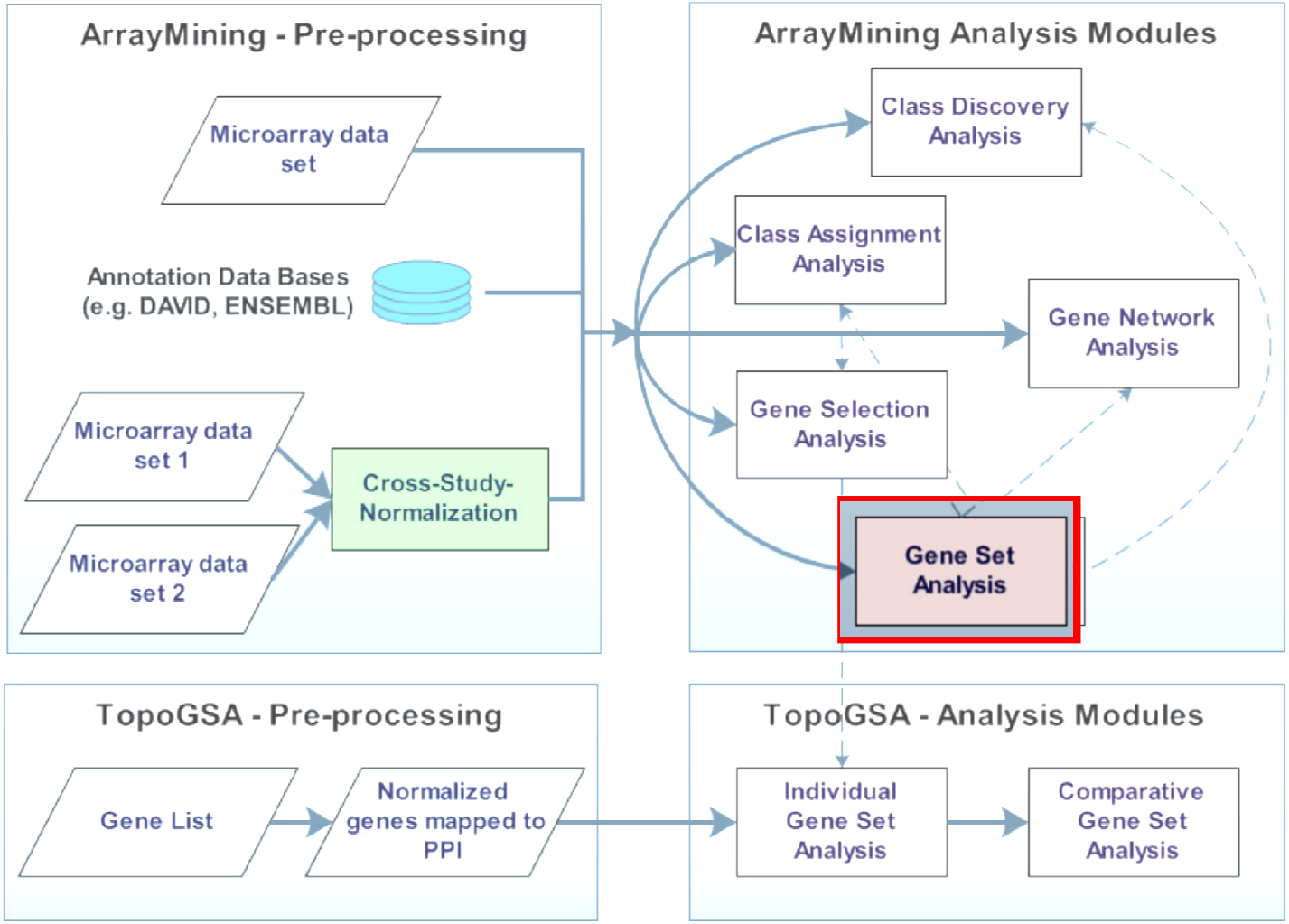
Example application: QMC breast cancer dataset

- Separate sub-classes in 84 luminal samples with consensus clustering
- Input algorithms: k-Means, SOM, SOTA, PAM, HCL, DIANA, HYBRID-HCL





Methods overview: ArrayMining & TopoGSA



Gene set analysis – Motivation:

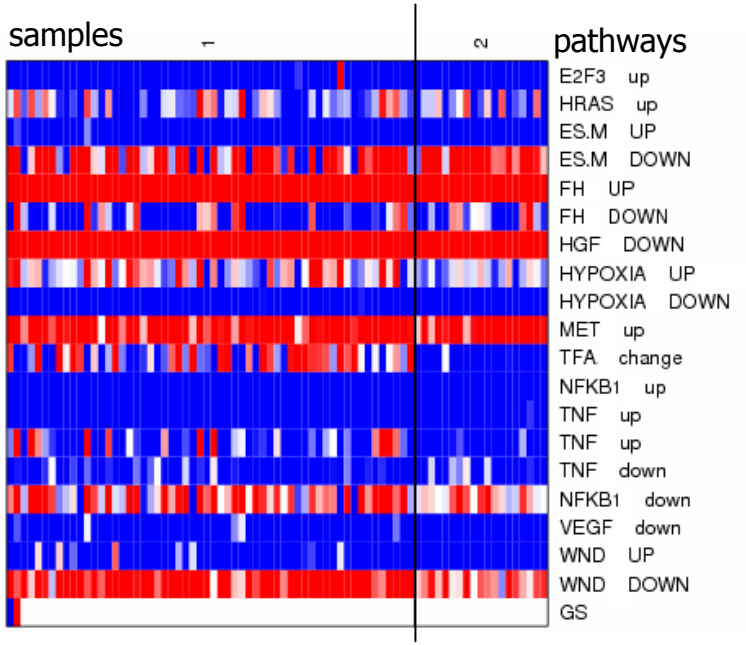
- Measurements for a single gene are often unreliable
- Similar genes might contain complementary information
- We want to integrate functional annotation data

→ **Gene Set Analysis (GSA)**:

1) Identify sets of functionally similar genes (GO, KEGG, etc.)

2) Summarize gene sets to „Meta"-genes (PCA, MDS, etc.)

3) Apply statistical analysis

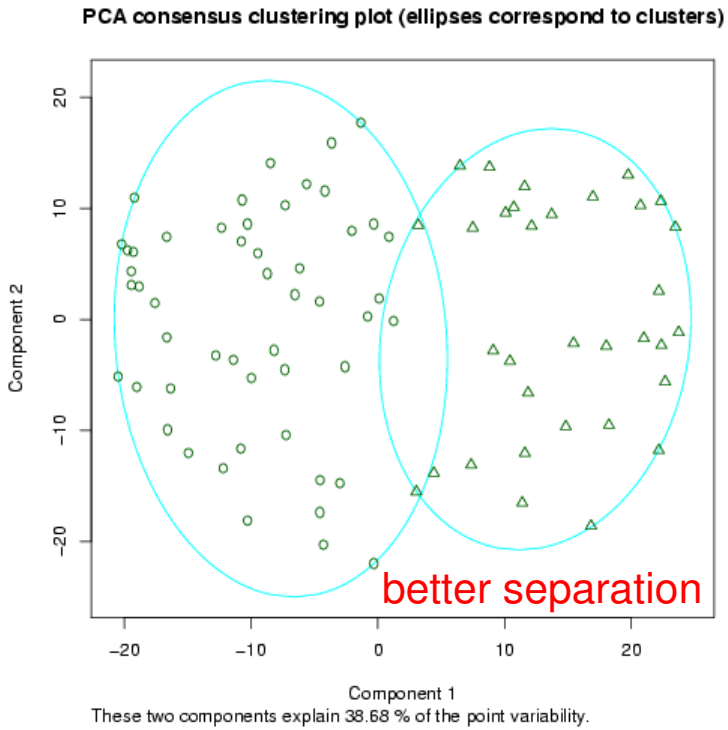


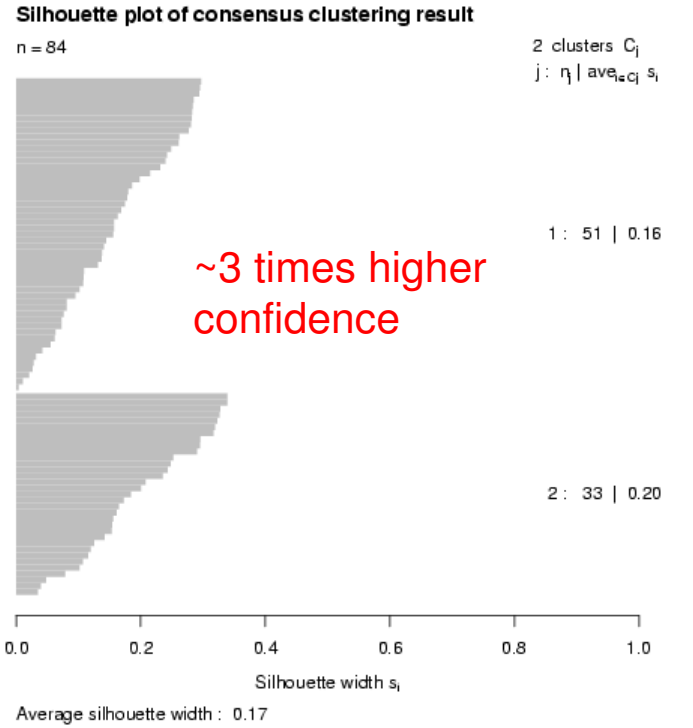
(example: Van Andel institute cancer gene sets)

Consensus clustering: example (2)

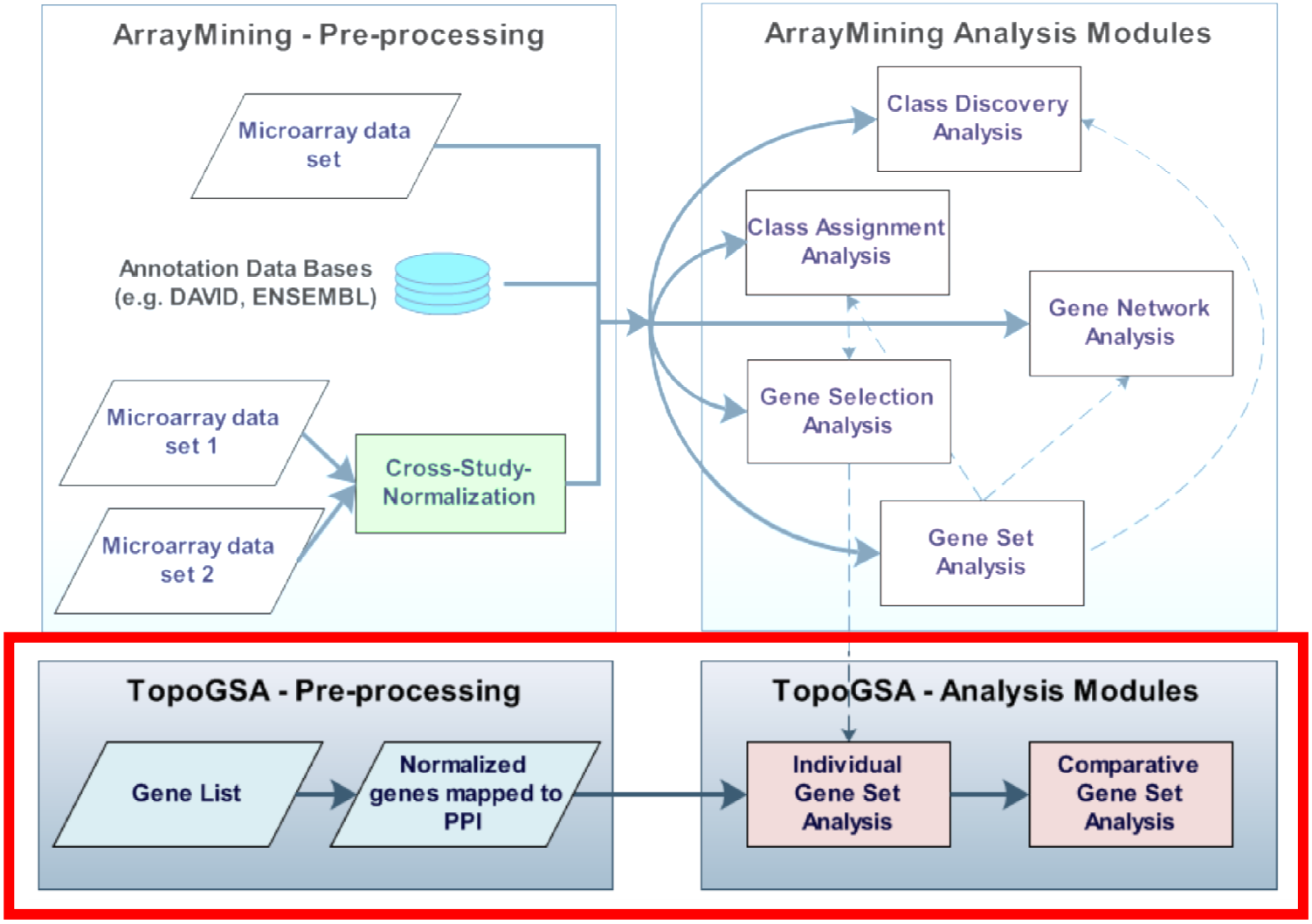
Combine consensus clustering with gene set analysis

- Map genes onto Gene Ontology (GO), reduce dimensionality (MDS)
- Apply same consensus clustering as before on GO-based „meta-genes“



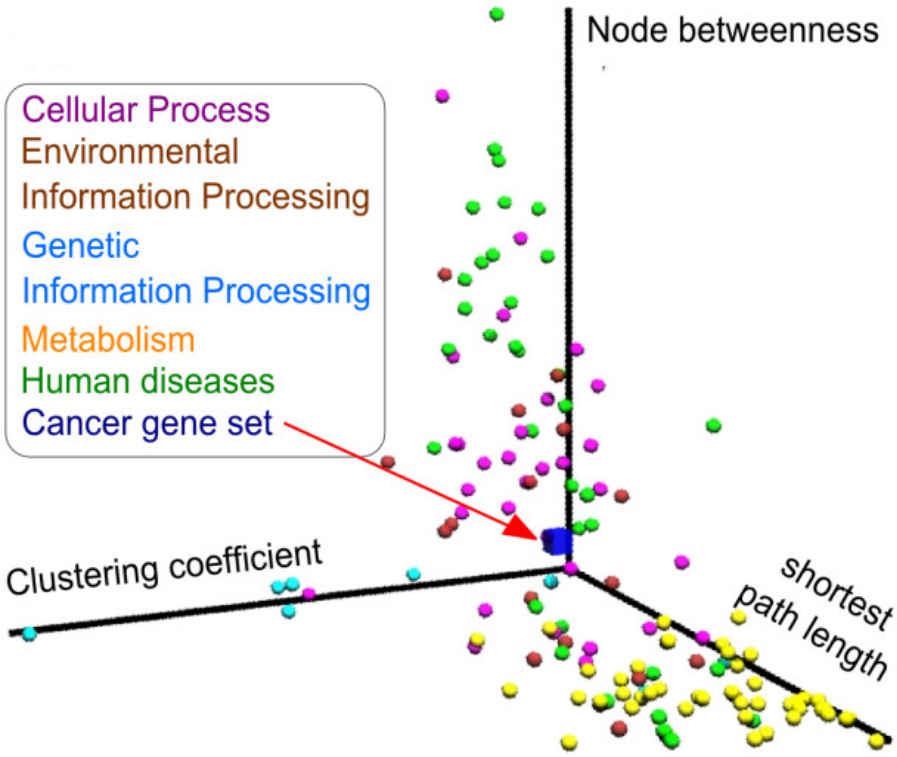


Methods overview: ArrayMining & TopoGSA



TopoGSA: Network topological analysis of gene sets

www.infobiotics.net/TopoGSA



What is TopoGSA?

TopoGSA is a web-application mapping gene sets onto a comprehensive human protein interaction network and analysing their network topological properties.

Two types of analysis:

1. Compare genes within a gene set: e.g. up- vs. down-regulated genes

2. Compare a gene set against a database of known gene sets (e.g. KEGG, BioCarta, GO)

TopoGSA - Methods

TopoGSA computes the following topological properties for an uploaded geneset and matched-size random gene sets:

- the **degree** of each node in the gene set

- the **local clustering coefficient C_i** for each node v_i in the gene set:

$$C_i = \frac{2|e_{jk}|}{k_i(k_i-1)} : v_j, v_k \in v_j : e_{ji} \in Ee_{ij} \in E, e_{jk} \in E$$

  where ki is the degree of v_i and e_{jk} is the edge between v_j and v_k

- the **shortest path length** between pairs of nodes v_i and v_j in the gene set

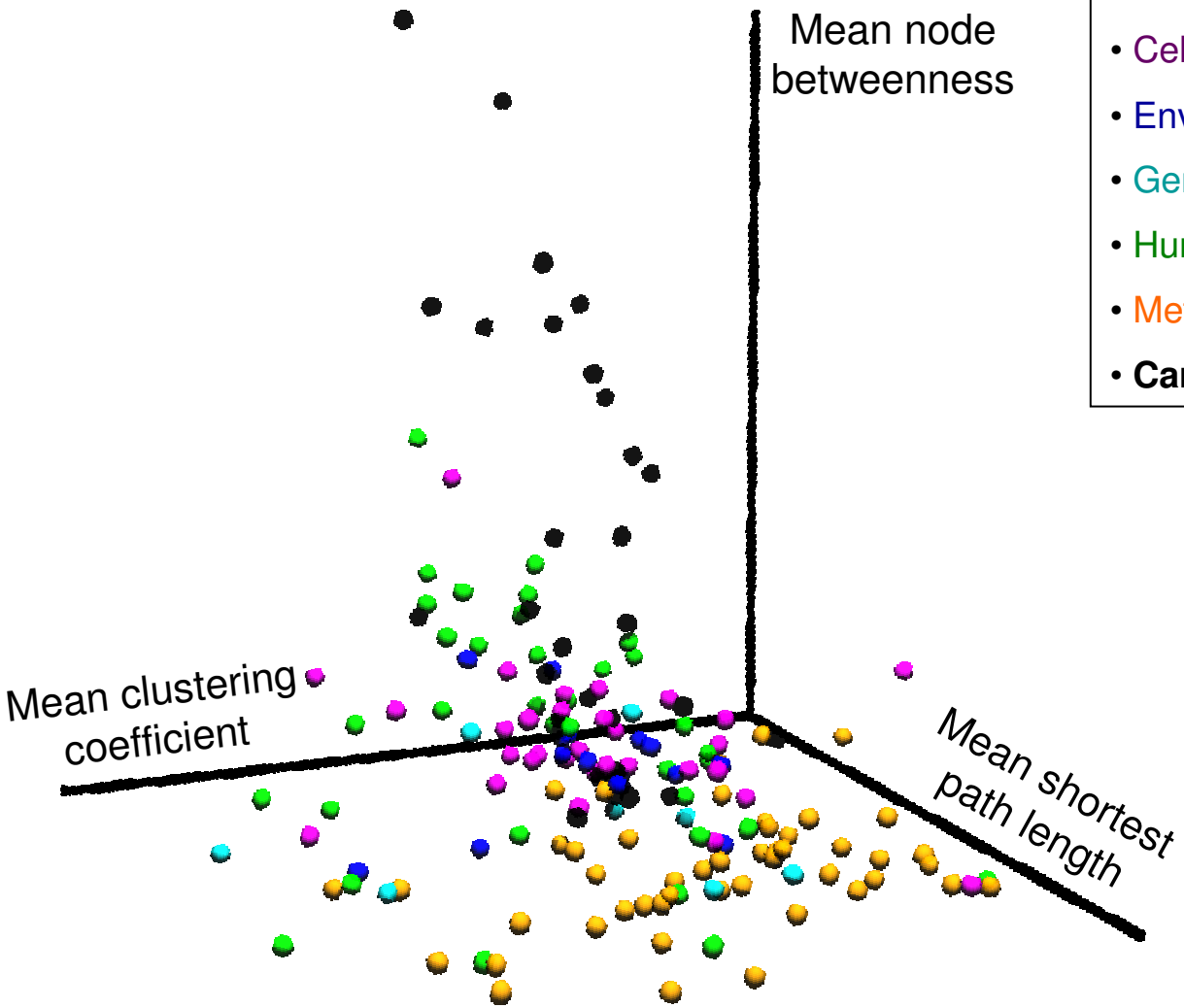
- the **node betweenness B(v)** for each node v in the gene set:

$$B(v) = \sum_{s \neq v, s \neq t, v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

  here $\sigma_{st}(v)$ is the number of shortest paths from s to t passing through v

- the **eigenvector centrality** for each node in the gene set

KEGG-BRITE pathway colouring



LEGEND:

- Cellular processes
- Environmental information processing
- Genetic information processing
- Human diseases
- Metabolism
- **Cancer genes**

General results:

- Metabolic pathways have high shortest path lenghts and low betweenness
- Disease pathways and cancer gene sets tend to have high betweenness and small shortest path lenghts

ArrayMining → TopoGSA

Send selected genes from ArrayMining to TopoGSA:

- Results of **within-gene-set comparison**:

  Estrogen receptor 1 gene and **apoptosis regulator Bcl2**, both up-regulated in luminal samples, have outstanding network topological properties (higher betweenness, higher degree, higher centrality) in comparison to other genes.

- Results of **comparison against reference databases**:

  - Metabolic **KEGG** pathways are most similar to the uploaded gene set in terms of network topological properties.

  - Most similar **BioCarta** pathways: Cytokine, differentiation and inflammatory pathways.

Conclusions / Outlook

- Combining algorithms in a sequential and/or parallel fashion can provide performance improvements and new biological insights
- Microarray and gene set analysis tasks can be interlinked flexibly in an (almost) completely automated process
- New analysis types like network-based topology analysis and co-expression analysis complement existing tools
- For further details: See our publications in BMC Bioinformatics (Glaab et al., 2009) and Bioinformatics (Glaab et al., 2010)

References

1. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *TopoGSA: network topological gene set analysis*, Bioinformatics, 26(9):1271-1272, 2010

2. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *Extending pathways and processes using molecular interaction networks to analyse cancer genome data*, BMC Bioinformatics, 11(1), 597, 2010

3. E. Glaab, J. M. Garibaldi and N. Krasnogor. *ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization*, BMC Bioinformatics,10:358, 2009

4. E. Glaab, J. M. Garibaldi, N. Krasnogor. *Learning pathway-based decision rules to classify microarray cancer samples*, German Conference on Bioinformatics 2010, Lecture Notes in Informatics (LNI), 173, 123-134

5. E. Glaab, J. M. Garibaldi and N. Krasnogor. *VRMLGen: An R-package for 3D Data Visualization on the Web*, Journal of Statistical Software, 36(8),1-18, 2010

6. H. O. Habashy, D. G. Powe, E. Glaab, N. Krasnogor, J. M. Garibaldi, E. A. Rakha, G. Ball, A. R. Green, C. Caldas, I. O. Ellis. *RERG (Ras-related and oestrogen-regulated growth-inhibitor) expression in breast cancer: A marker of ER-positive luminal-like subtype, Breast Cancer Research and Treatment*, 2010 (Epub ahead of print)