

Virtual Biology Group

Ronnie Alves (postdoc)

Richard Christen (group Leader)

Olivier Croce (postdoc)

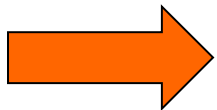
Julien Gardes (PhD student)

Claude Pasquier (researcher CNRS)

Medja Zermat (predoc)

Current biology is an information-intensive science

- High throughput technologies are generating large amounts of data,
 - distributed across many data sources
 - Web pages
 - Flat files
 - Databases (public & proprietary)
 - Spreadsheets
 - with high degree of semantic heterogeneity
 - and different levels of quality.



Efficient analyses require integrating such data, extracting information and interpreting it in order to generate biological knowledge

Virtual Biology Lab



Our research

to develop and use computational techniques to address biological problems

Biological researches

- Biodiversity
 - Study of microbial diversity
- Genomics
 - Identification of human specific genes
- Transcriptomics
 - Integration and analysis of transcriptomic data

Computer science researches

- Information integration
- Knowledge modeling
- Visualization
- Data mining

Virtual Biology Lab



Our research

to develop and use computational techniques to address biological problems

Biological researches

- Biodiversity
 - Study of microbial diversity
- Genomics
 - Identification of human specific genes
- Transcriptomics
 - Integration and analysis of transcriptomic data

Computer science researches

- Information integration
- Knowledge modeling
- Visualization
- Data mining

GenMiner:

Mining Informative Association Rules from Genomic Data

Ricardo Martinez¹, **Claude Pasquier**² and Nicolas Pasquier¹

¹ I3S - Laboratory of Computer Science, Signals and Systems

² IDBC - Institute of Developmental Biology and Cancer

University of Nice – Sophia Antipolis - France

Mail: **claude.pasquier@unice.fr**

Data Mining

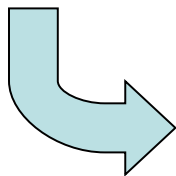
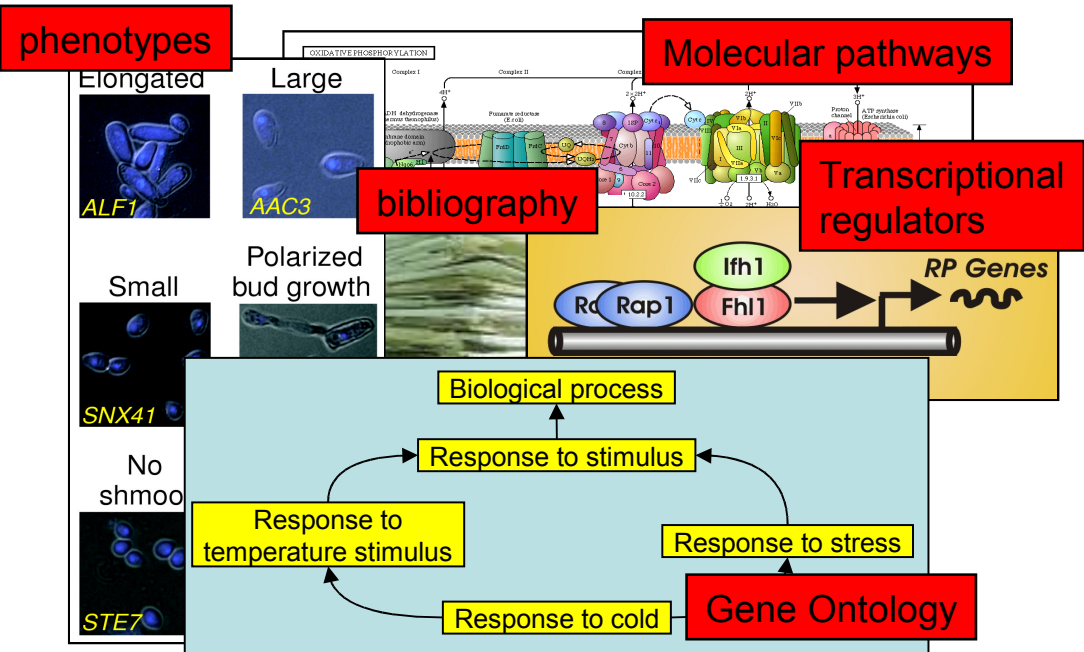
- Data Mining is the use of computers to help make sense of information.
- It uses special techniques to find hidden patterns in data
- Notable applications in:
 - Marketing
 - market basket analysis
 - customer profiling
 - Banking
 - credit risk analysis
 - fraud detection
 - Telecommunications
 - fraud detection
 - network faults isolation
- Data mining in biology is an emerging research field

Data mining for transcriptomic studies

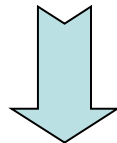
Transcriptomic data (measures)



Background biological data



Data mining



Information

Several data mining approaches

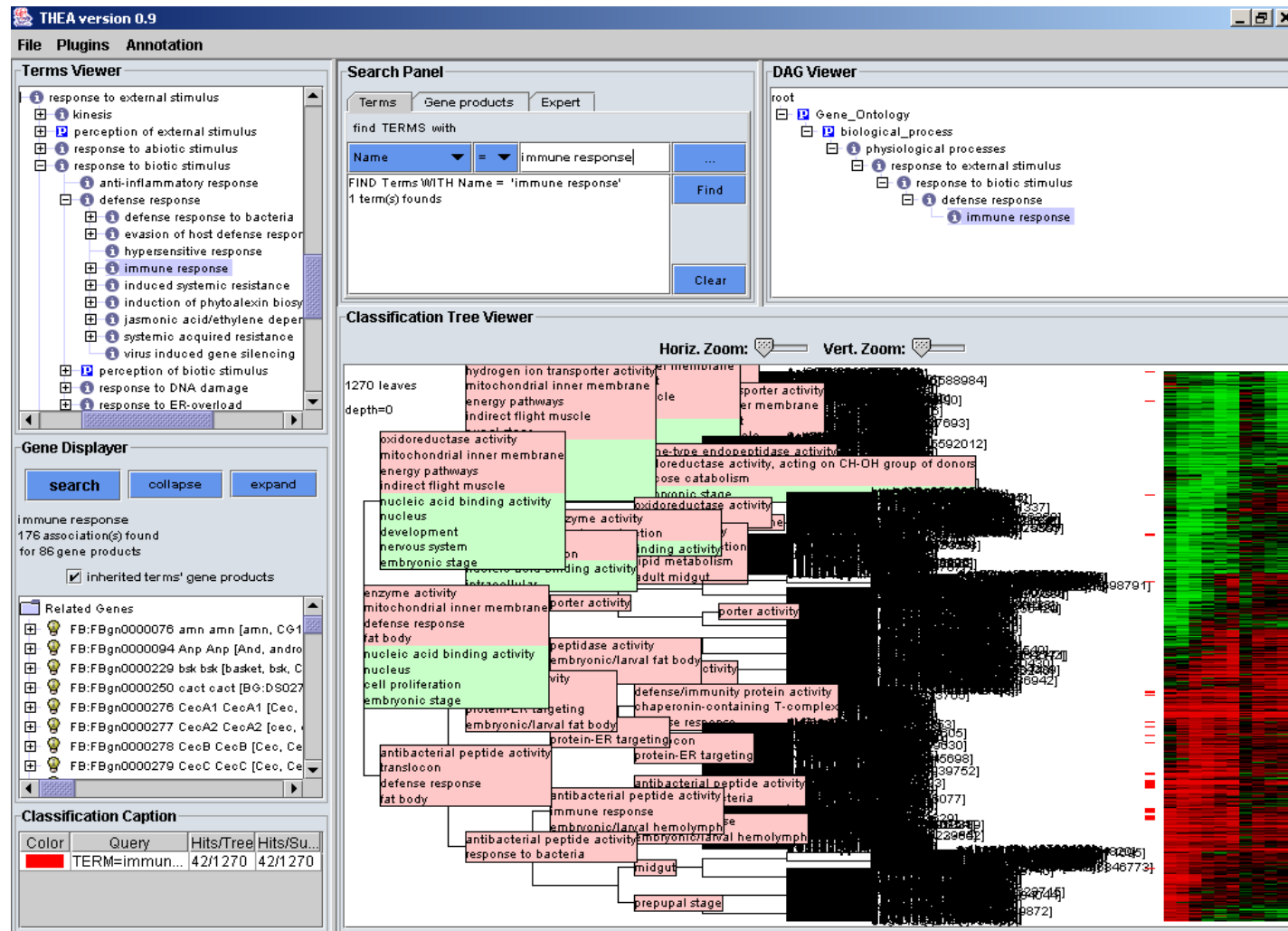
- Expression-based (Eisen et al., 1998)
 - Groups of co-expressed genes are built (clustering)
 - Biological data are integrated
- Knowledge-based (Mootha et al., 2003)
 - Groups of co-annotated genes are built
 - Gene expressions are integrated
- Co-clustering (Hanish et al., 2002)
 - Groups of co-expressed and co-annotated genes are built

Eisen MB, Spellman PT, Brown PO and Botstein D. Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc Natl Acad Sci U S A 95, 14863-8, (1998)

Mootha VK et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34, 267 - 273 (2003)

Hanisch D, Zien A, Zimmer R, Lengauer T. Co-clustering of biological networks and gene expression data. Bioinformatics 18, supp1 (2002)

Expression based approach



Pasquier C, Girardot F, Jevardat de Fombelle K and Christen R. THEA: ontology-driven analysis of microarray data. *Bioinformatics* 20(16):2636-43 (2004)

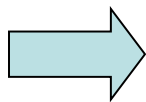
Drawbacks of expression based approach

Over-expressed genes that are highlighted in this experiment are not clustered together

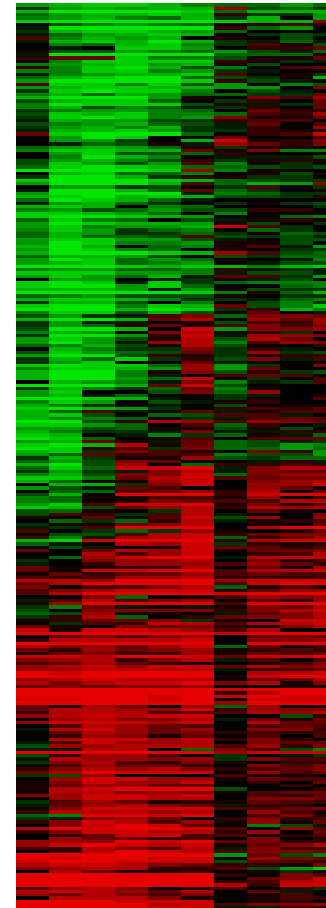
The result only shows relationships (not directed) between sets of co-expressed genes and sets of co-annotated genes.

One might be interested by results expressing directed relationships on the form:

- Genes that are over-expressed in exp1 and in exp4 are related to the same function
- Genes belonging to a given pathway are under-expressed in exp3 and in exp4
- Genes that are under-expressed in exp2 are also under-expressed in exp4 and over-expressed in exp5



Association Rule discovery



Association Rules

- A directed relationship between two sets of attribute values
- Attributes represent discrete values
- Typical application: market basket analysis
 - Example of sales data:
 - Buy:cereals ^ buy:sugar → buy:milk [support = 10%, confidence = 60%]
- Support: weight/scope of the rule
 - Proportion of objects (tuples) containing all items
 - 10% of all customers have bought all three items
 - Useful rule : support $\geq$ minsupport
- Confidence = precision/reliability of the rule
 - Proportion of objects containing the consequent among those containing the antecedent
 - 60% of customers having bought cereals and sugar also have bought milk
 - Significant rules : confidence $\leq$ minconfidence

ARD applied to gene expression analysis

- Tuzhilin 2002, Creighton 2003, Georgi 2005
 - [expression profiles]⁺ → [expression profiles]⁺
 - gene1[↑] → gene2[↑], gene3[↓]
- Carmona-saez 2006
 - [annotation]⁺ → [expression profile]⁺
 - e.g. annotation1 → experiment1[↓], experiment2[↑]

Tuzhilin A and Adomavicius G. Handling Very Large Numbers of Association Rules in the Analysis of Microarray Data. Proceedings ACM SIGKDD International Conference (2002)

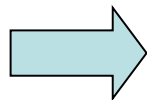
Creighton C and Hanash S. Mining gene expression databases for association rules. Bioinformatics 19:79-86 (2003)

Georgii E, Richter L, Rückert U and Kramer S. Analyzing microarray data using quantitative association rules. Bioinformatics 21:123-9 (2005)

Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM and Pascual-Montano A. Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics. 7;7:54 (2006)

Limitations of ARD applied to transcriptomics

- Efficient approach for weakly correlated and sparse data
- Biological data are highly correlated data
- Analysis of correlated data constitute a challenge
 - Problem of execution time
 - Several hours in most cases (occasionally several days)
 - Problem of discretization of gene expression data
 - Problem of rules utility and relevance :
 - Several tens of thousands of rules (occasionally millions)
 - Among these rules many are redundant (same information)



GenMiner

Martinez R, Pasquier C, Pasquier N. GenMiner: Mining Informative Association Rules from Genomic Data. Proceedings of the IEEE BIBM international conference on Bioinformatics and Biomedecine (BIBM'2007), november 2-4, 2007.

Martinez R., Pasquier C. and Pasquier N. GenMiner: Mining non-redundant association rules from integrated gene expression data and annotations. Bioinformatics Advance Access published on September 17, 2008

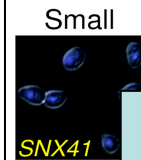
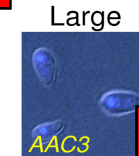
GenMiner Framework

Transcriptomic data (measures)



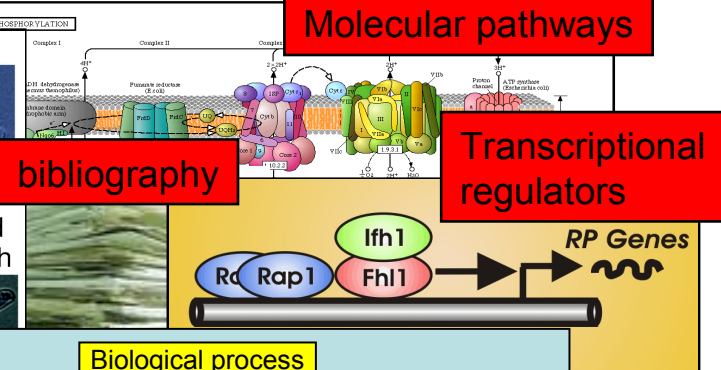
Background biological data

phenotypes



bibliography

Molecular pathways



Transcriptional regulators

Biological process

Response to stimulus

Response to temperature stimulus

Response to stress

Response to cold

Gene Ontology

discretization

NorDi

association rules extraction

GenMiner

Close

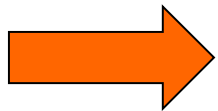
minimal association rules

Data analyzed by GenMiner

- Gene expression measures
 - Eisen dataset 2465 yeast genes X 79 experimental conditions (cell cycle, sporulation, temperature shock and diauxic shift experiments)
- Gene annotations
 - 581 bibliographic annotation from SGD's manually curated papers
 - 14 annotations relative to metabolic pathways from KEGG
 - 14 phenotype data from SGD
 - 25 transcriptional regulators
 - 24 Gene Ontology (GO) annotations using GoSlim
- Analyzed dataset
 - A matrix of 2465 lines (genes) and 737 columns (annotations and expression levels discretized by NorDi).

Associations annotations → expressions

Rule	Antecedent	Consequent	Supp. (#)	Conf. (%)
1	go:0006412 (translation) go:0005840 (ribosome)	heat3↓	103	51
2	go:0042254 (ribosome biogenesis and assembly) go:0005840 (ribosome) go:0005198 (structural molecule activity)	heat4↓	15	52
3	path:sce03010 (ribosome pathway)	heat4↓	69	53
4	tr:RAP1 tr:FHL1	heat3↓	71	62
5	pmid:5542014, pmid:9649613, pmid:3533916	heat3↓	12	100



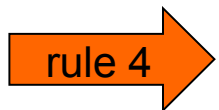
general reduction of protein synthesis, ribosomal organization and cell maintenance following a heat shock.

Associations expressions → annotations

Rule	Antecedent	Consequent	Supp. (#)	Conf. (%)
1	heat3↓, heat4↓, heat5↓	Go:0006412 (translation)	35	88
2	heat2↓	Go:0006996 (organelle organization and biogenesis)	41	69
3	heat2↓	Go:0042254 (ribosome biogenesis and assembly)	39	66
4	heat2↑, heat3↑, heat4↑	Go:0006950 (response to stress)	15	52



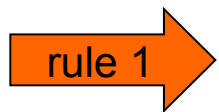
genes under-expressed during the heat shock experiment
are involved in protein synthesis, cellular organization and
ribosomal organization



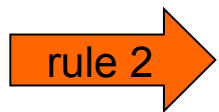
genes over-expressed are involved in stress response

Associations expressions → expressions

Rule	Antecedent	Consequent	Supp. (#)	Conf. (%)
1	heat4↓	heat3↓	83	79
2	cold4↓	heat3↓	66	68



79 % of the genes that are under-expressed at the time point 4 of the heat shock experiment were already under-expressed at the time point 3



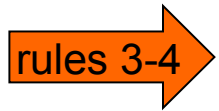
a set of genes are under-expressed after both a heat shock and a cold shock

Associations annotations → annotations

Rule	Antecedent	Consequent	Supp. (#)	Conf. (%)
1	Path:sce00190 (purine metabolism)	Go:0005737 (cytoplasm)	52	96
2	Pmid:16155567	Phenot:inviable	96	94
3	Tr:FHL1	Tr:RAP1	114	86
4	Tr:RAP1	Tr:FHL1	114	61



Revealing of possible links between annotations from different sources



RAP1 binding is essential for the recruitment of FHL1
(Identification of portions of gene regulatory network)

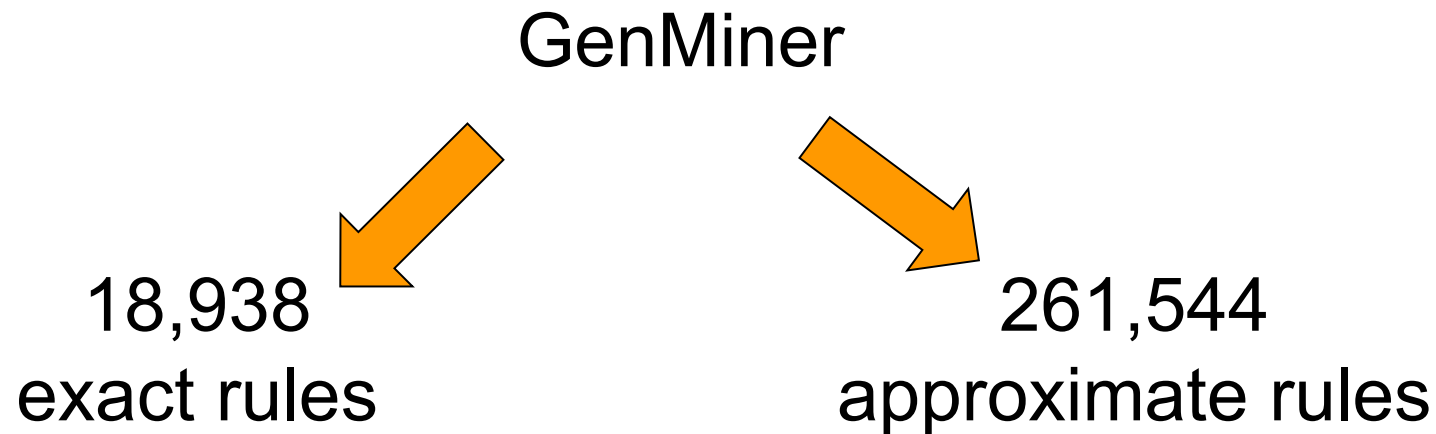
PMID:16155567: Davierwala et al. The synthetic genetic interaction spectrum of essential genes. Nat Genet. 37(10):1147-52 (2005)

Conclusion

- GenMiner is a powerful tool for finding rules hidden into large sets of biological data
 - It allows the mining of gene expressions
 - continuous values are discretized by NorDi
 - It facilitates the analysis of rules
 - Close generates minimal sets of non-redundant rules
 - No a-priori in the structure of the rules
- However, finding meaningful rules is difficult because we lack biological expertise
 - Rules with high support and high confidence are not necessarily “interesting rules”
 - We can only check if information described in research papers is highlighted by our tools
- We are looking for collaborations
 - You can help us to apply our algorithms to new datasets
 - We can help you to highlight important information hidden in your data

Processing

- min support = 0.5% (at least 12 lines)
- min confidence = 50%
- execution time: 16 minutes



Results

- 73% of the rules (206190 out of 280,482) are difficult to interpret
 - Tr:FHL1, spo2↓ → go:0005737 (cytoplasm), heat3↓
supp=50 conf=0,91
- 2% of the rules (6,209 out of 280,482) are relative to the GO hierarchy
 - Go:0005739 (mitochondrion) → Go:0005737 (cytoplasm)
supp=503 conf=1

Results

- The remaining rules correspond to these patterns:
 - 31179 rules annotations → expressions (11%)
 - 13128 rules expressions → annotations (5%)
 - 13943 rules expressions → expressions (5%)
 - 9833 rules annotations → annotations (4%)

Number of rules extracted from typical dataset

- Datasets:
 - T10I4D100K: simulated market basket data (weakly correlated data)
 - C73D10K: sample from U.S. Census Bureau (highly correlated data)

Dataset	Min support	Min confidence	Nb rules
T10I4D100K	0.5%	100%	0
		70%	20,419
		50%	21,686
C73D10K	90%	100%	52,035
		95%	1,606,726
		80%	2,053,936

Number of informative rules

- Datasets:
 - T10I4D100K: simulated market basket data (weakly correlated data)
 - C73D10K: sample from U.S. Census Bureau (highly correlated data)

Dataset	Min support	Min confidence	Nb rules	Inf rules
T10I4D100K	0.5%	100%	0	0
		70%	20,419	4,004
		50%	21,686	4,191
C73D10K	90%	100%	52,035	1,369
		95%	1,606,726	5,680
		80%	2,053,936	5,718