UMG

Dubium sapientiae initium

Efficient preprocessing and analysis of omics data:

Experiences at University Magna Graecia of Catanzaro

Mario Cannataro

Bioinformatics Laboratory &
Data Analytics Research Center,
Dep. of Medical and Surgical Sciences
University "Magna Græcia" of Catanzaro, Italy

cannataro@unicz.it

CGW 2017, Krakow, Poland – 23-25 October, 2017

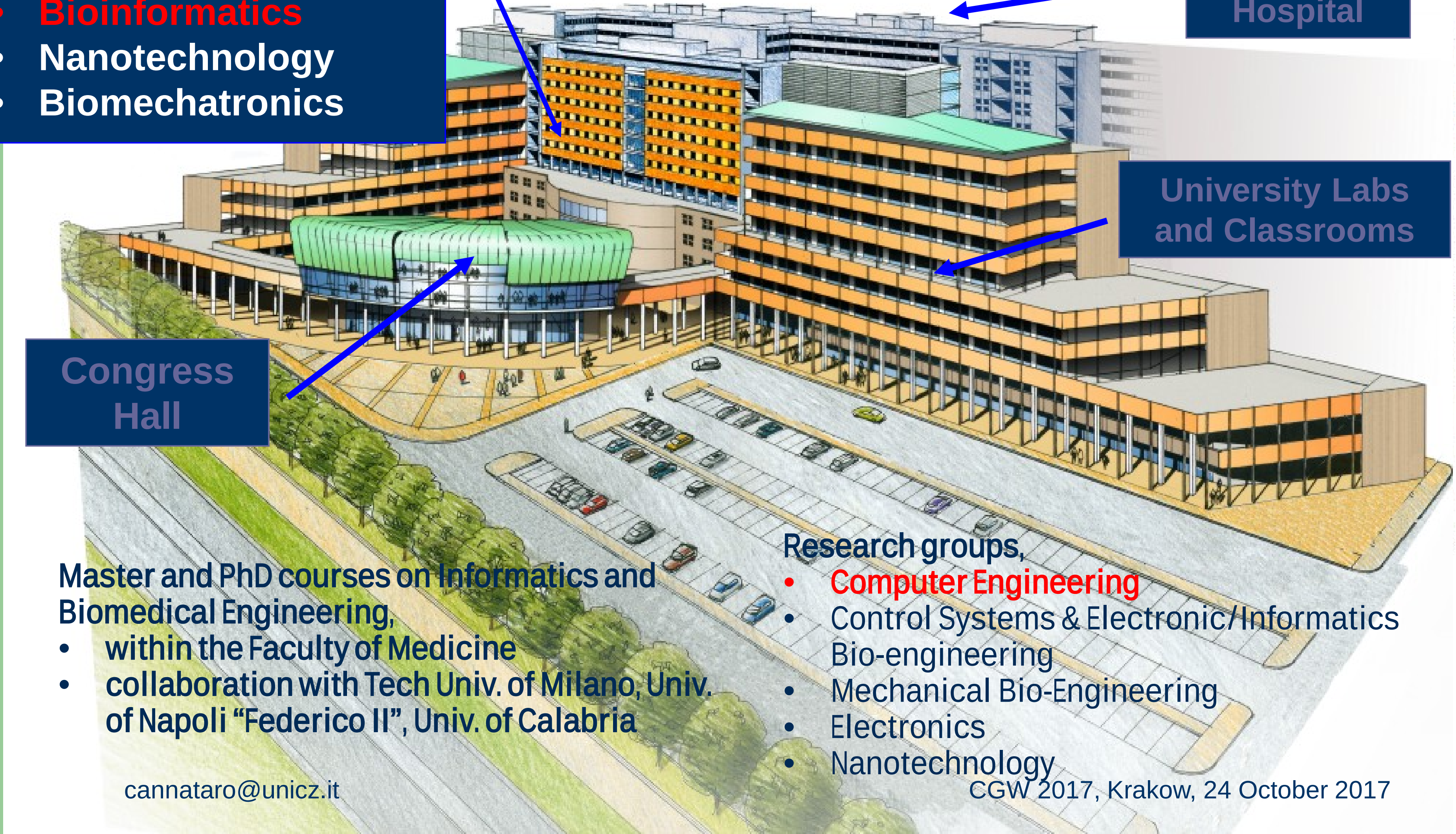
Brief History

- 1998: Format start of University of Catanzaro
- 2001/2002: Start of a Inter-University **Bachelor on Informatics and biomedical engineering**
- 2003: **University Campus** in its initial stage (right)
- 2004: Start of the interdisciplinary **PhD Program on Informatics and biomedical engineering**
- 2017: **University Campus** now (bottom)

Core Facilities

- Genomics
- MicroArray
- Proteomics
- Mass Spectrometry
- Bioinformatics
- Nanotechnology
- Biomechatronics

University of Catanzaro, Campus "Salvatore Venuta", Medicine and Bioscience Buildings



University Hospital

University Labs and Classrooms

Congress Hall

Master and PhD courses on Informatics and Biomedical Engineering,

- within the Faculty of Medicine
- collaboration with Tech Univ. of Milano, Univ. of Napoli "Federico II", Univ. of Calabria

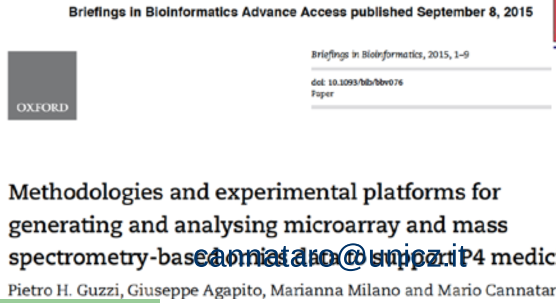
Research groups,

- Computer Engineering
- Control Systems & Electronic/Informatics Bio-engineering
- Mechanical Bio-Engineering
- Electronics
- Nanotechnology

P4 Medicine

- **Predictive, preventive, personalized and participatory (P4) medicine** is an emerging medical model that is based on the customization of all medical aspects (i.e. practices, drugs, decisions) of the individual patient
  - **Predictive medicine**: early diagnosis of a large set of diseases through regular scanning of the omic data (e.g. biomarker discovery)
  - **Preventive medicine**: prevent the development of some diseases before the appearance of symptoms through the monitoring of ‘wellness’, i.e. the behaviour of patients avoiding possible unhealthy practices
  - **Personalized medicine (Precision medicine)** is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. **Pharmacogenomics has a key role.**
    - January 20, 2015, President Obama announced the Precision Medicine Initiative® (PMI) - NIH
  - **Participatory medicine**: the active involvement of patients in the medical processes (e.g. patient-driven networks).
    - ‘Participatory Medicine is a model of cooperative health care that seeks to achieve active involvement by patients, professionals, caregivers, and others across the continuum of care on all issues related to an individual’s health.’ (Society for Participatory Medicine)

P4 MEDICINE institute



Briefings in Bioinformatics Advance Access published September 8, 2015

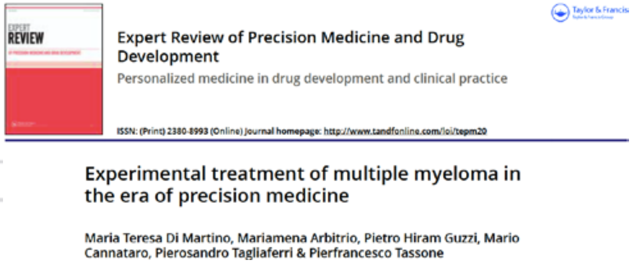
OXFORD

Briefings in Bioinformatics, 2015, 1–9

doi: 10.1093/bib/bbv076
Paper

Methodologies and experimental platforms for generating and analysing microarray and mass spectrometry-based omics data to support P4 medicine

Pietro H. Guzzi, Giuseppe Agapito, Marianna Milano and Mario Cannataro



Taylor & Francis

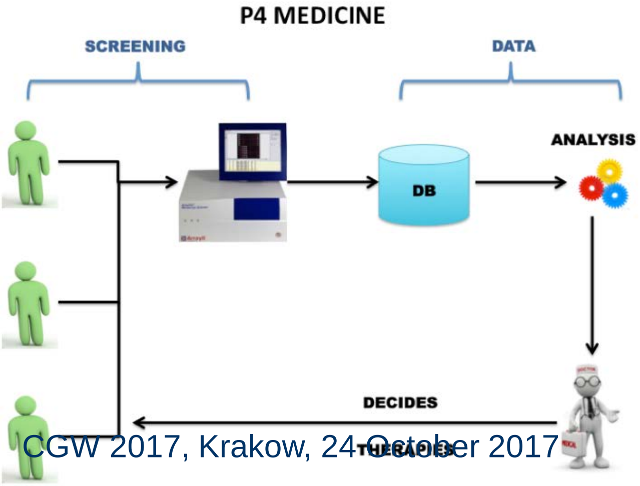
Expert Review of Precision Medicine and Drug Development

Personalized medicine in drug development and clinical practice

ISSN: (Print) 2380-8993 (Online) Journal homepage: http://www.tandfonline.com/loi/tepm20

Experimental treatment of multiple myeloma in the era of precision medicine

Maria Teresa Di Martino, Mariamena Arbitrio, Pietro Hiram Guzzi, Mario Cannataro, Pierosandro Tagliaferri & Pierfrancesco Tassone



OUTLINE

- Experiences at University of Catanzaro in high performance management, preprocessing and analysis of omics data
  - PART I: Genomics data
  - PART II: Proteomics data
  - PART III: Interactomics data
- Conclusions

Part I: Genomics Data

- Microarray-based Genomics Data
  - gene expression and genotyping data
- Automatic preprocessing of gene expression data
  - **micro-CS**: preprocessing and annotation of gene expression data
- Statistical analysis of genotyping data
  - **DMET-Analyzer**: preprocessing and analysis of DMET (Drug Metabolizing Enzymes and Transporters) SNP (Single Nucleotide Polymorphism) data
  - **coreSNP** (multicore implementation)
- Data Mining analysis of genotyping data
  - **DMET-Miner**: Association Rule Mining to extract from DMET data Association Rules able to correlate the contemporary presence of SNPs with patient’s conditions (e.g. TOX vs NOTOX)

Part I: Genomics Data

- Microarray-based Genomics Data
  - gene expression and genotyping data
- Automatic preprocessing of gene expression data
  - **micro-CS**: preprocessing and annotation of gene expression data
- Statistical analysis of genotyping data
  - **DMET-Analyzer**: preprocessing and analysis of DMET (Drug Metabolizing Enzymes and Transporters) SNP (Single Nucleotide Polymorphism) data
  - **coreSNP** (multicore implementation)
  - **cloud4SNP** (cloud-based implementation)
- Data Mining analysis of genotyping data
  - **DMET-Miner**: Association Rule Mining to extract from DMET data Association Rules able to correlate the contemporary presence of SNPs with patient's conditions (e.g. TOX vs NOTOX)

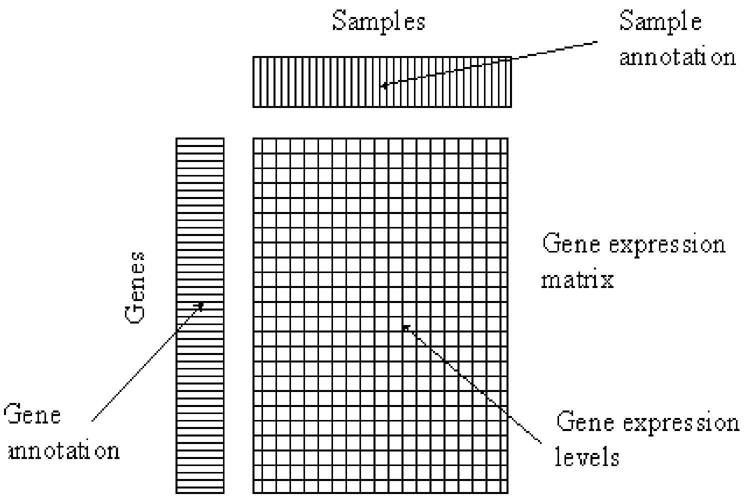
Gene Expression Microarray

- Sample Preparation
- Sample Deposition on Chip
- Ibridization
- Raw Data (Fluorescence Images)
- Preprocessing and Analysis

MeV MultiExperiment Viewer

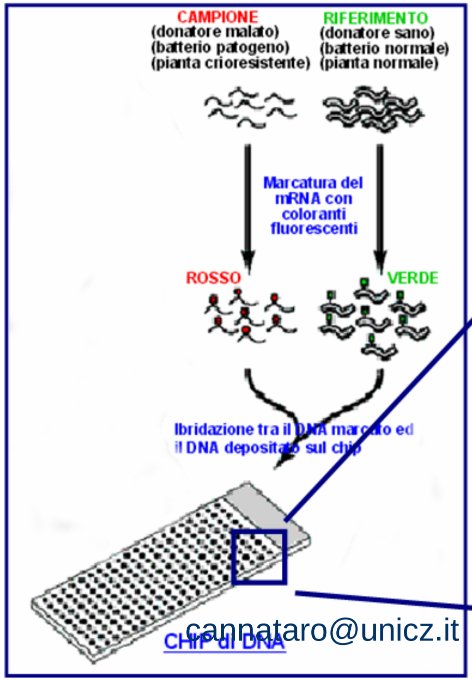
Analysis

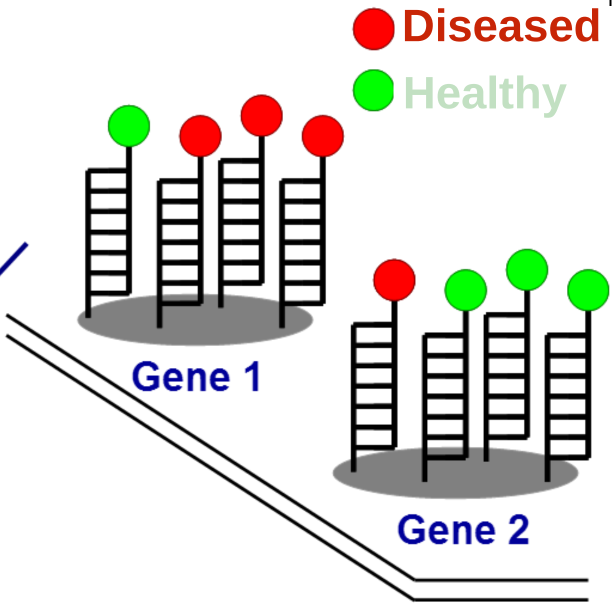
WEKA The University of Waikato

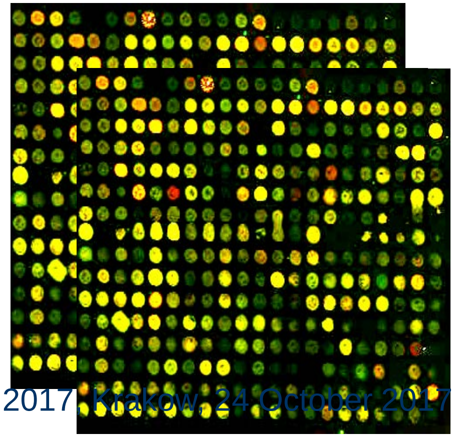


Summarization, Normalization, Annotation

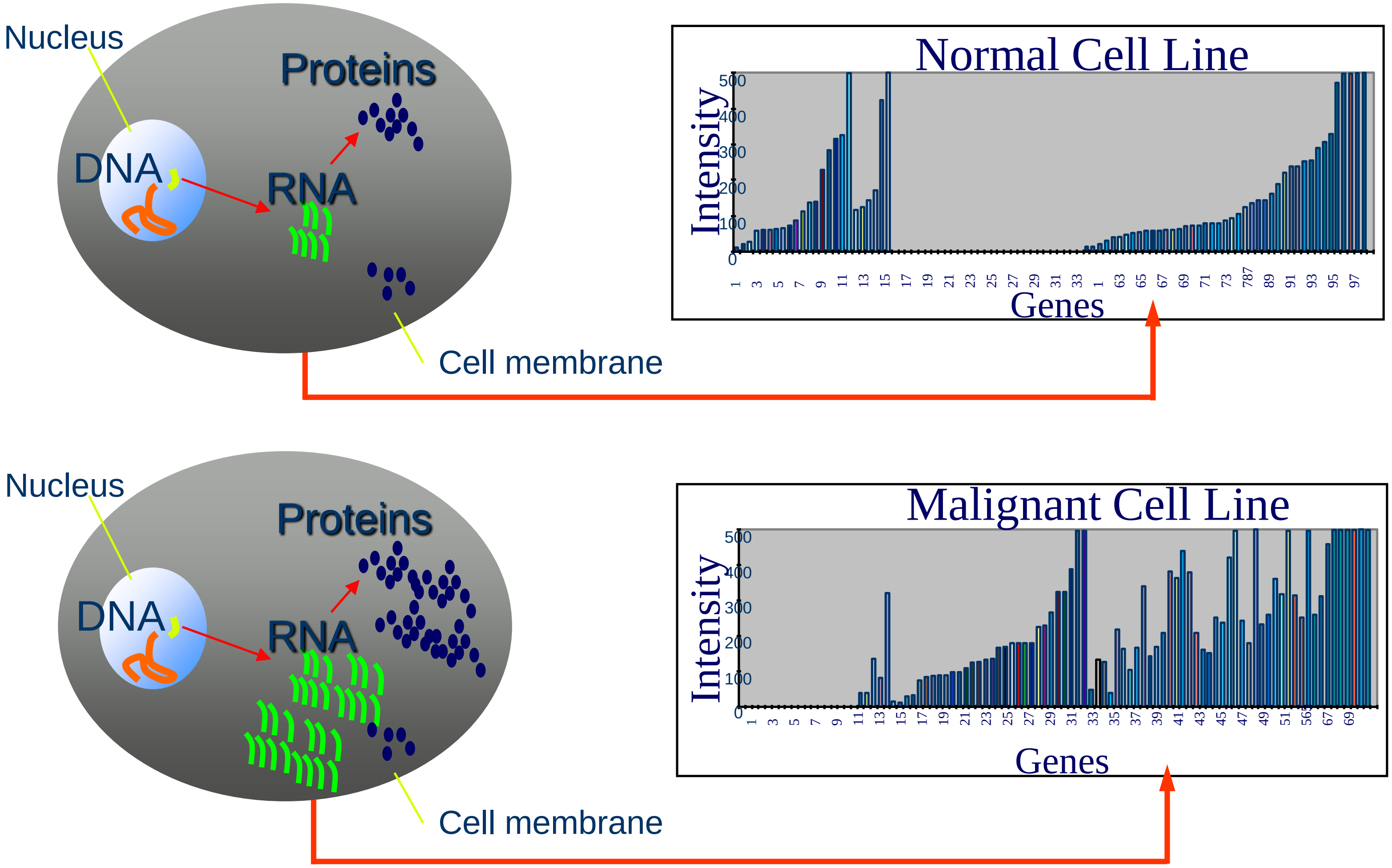
chip-specific libraries





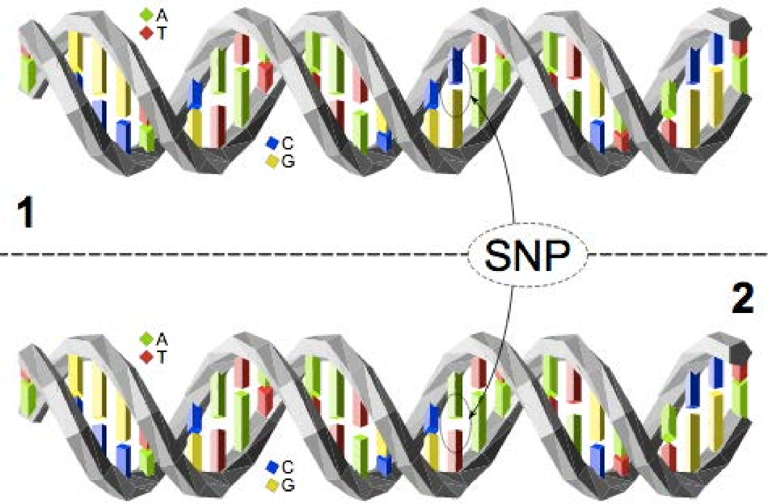


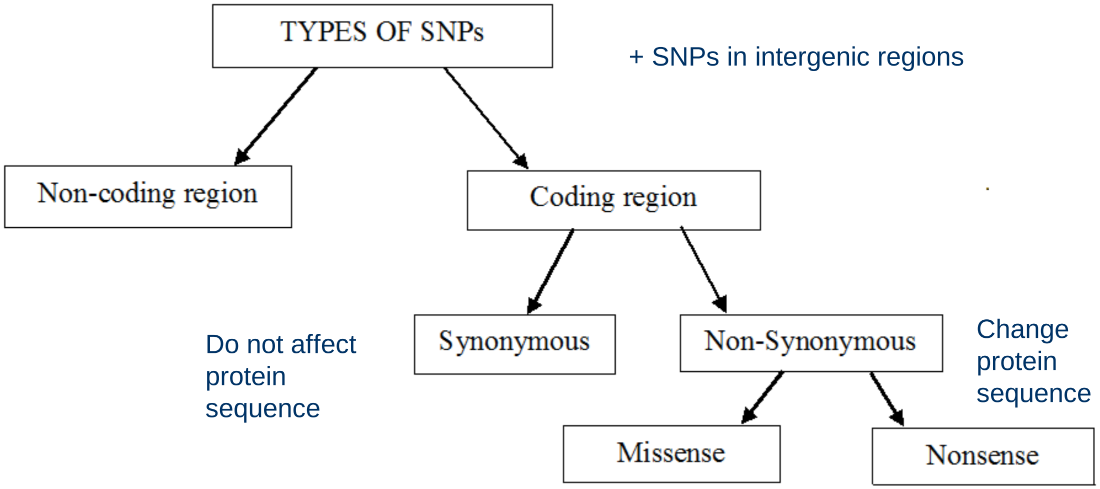
Gene expression = RNA “volume”



Single Nucleotide Polymorphisms (SNPs)

- A SNP is a variation in a single nucleotide that occurs at a specific position in the genome and is present in an appreciable percentage (e.g. >1%) within a population
- When in a specific base position, a base (e.g. C) appears in most individuals, but in a minority of individuals the position is occupied by a different base (e.g. A) then there is a SNP at that specific base position.





+ SNPs in intergenic regions

Do not affect protein sequence

Change protein sequence

Result in a different amino acid in the sequence protein

Result in a truncated, incomplete, and usually nonfunctional protein

Microarray-based Genomics data analysis

- We considered **gene expression** data and **SNP** (Single Nucleotide Polymorphism) **genotyping data** produced using microarrays
- We addressed two main problems:
  - The automation of the preprocessing pipeline of gene expression microarray data
    - **micro-CS**: preprocessing and annotation of gene expression data
  - The efficient (parallel) analysis of SNP genotyping microarray data
    - **DMET-Analyzer**: statistical analysis of DMET (Drug Metabolizing Enzymes and Transporters) SNP data
    - **coreSNP**: a parallel implementation of DMET-Analyzer
    - **DMET-Miner**: Association Rules correlating the presence of SNPs with patient’s conditions (e.g. TOX vs NOTOX)

Part I: Genomics Data

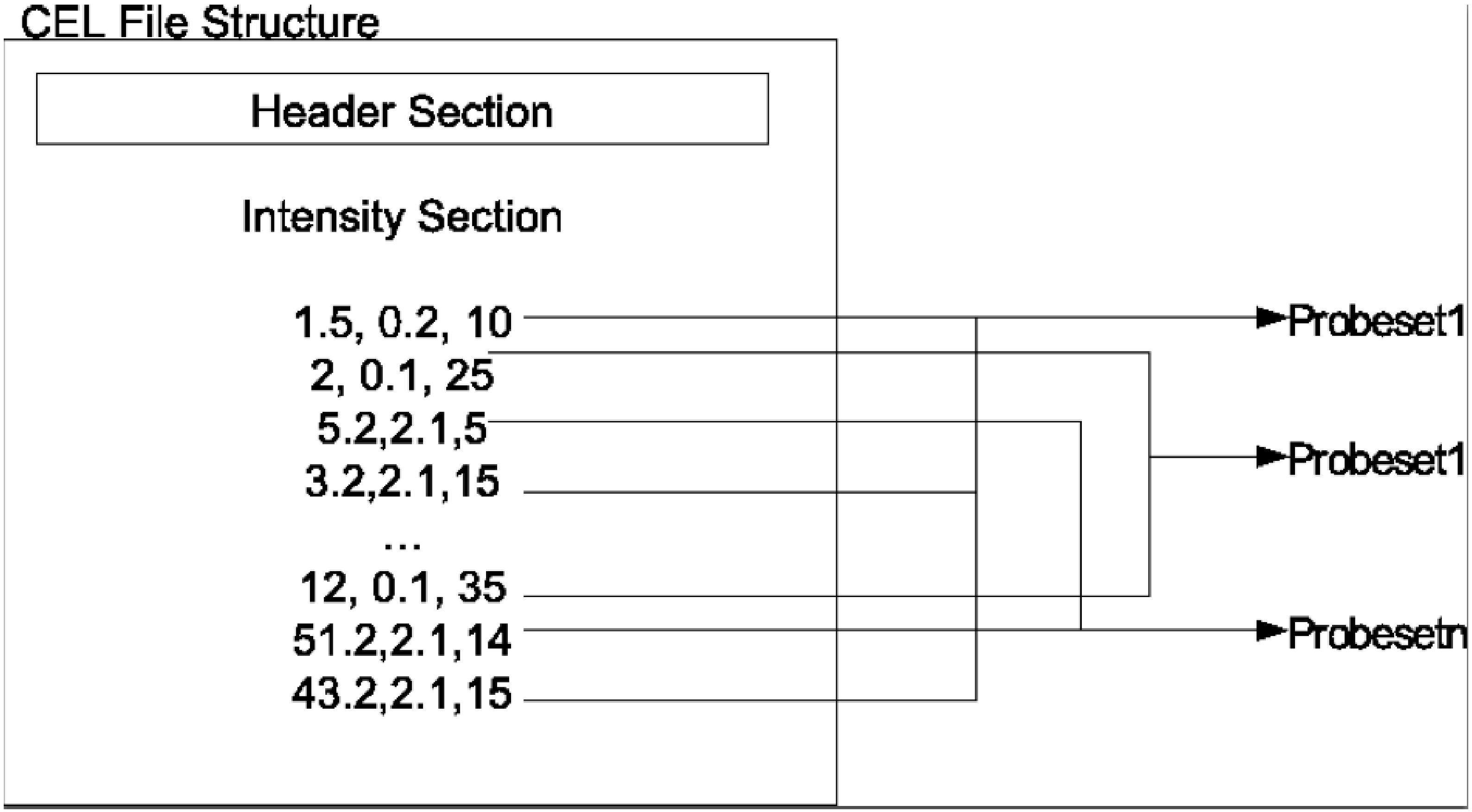
- Microarray-based Genomics Data
- Automatic preprocessing of gene expression data
  - **micro-CS**: preprocessing and annotation of gene expression data
- Statistical analysis of genotyping data
  - **DMET-Analyzer**: preprocessing and analysis of DMET (Drug Metabolizing Enzymes and Transporters) SNP (Single Nucleotide Polymorphism) data
  - **coreSNP** (multicore implementation)
- Data Mining analysis of genotyping data
  - **DMET-Miner**: Association Rule Mining to extract from DMET data Association Rules able to correlate the contemporary presence of SNPs with patient’s conditions (e.g. TOX vs NOTOX)

Workflow of analysis of gene expression data

- **(i) preprocessing**:
  - *Summarisation* recognizes the position of different genes in raw images, associating different regions of pixels to the unique gene that generated them.
  - *Normalisation* corrects the variation of gene expression in the same array due to experimental bias.
  - **Affymetrix Power Tools** (**APT**) is the command-line tool provided by Affymetrix for preprocessing. It uses some **chip-specific libraries**
- **(ii) annotation**: associates each gene to functional information, such as biological processes, and a set of cross reference DB identifiers
  - Annotation is usually done **using vendor-provided libraries** (e.g. Affymetrix)
- **(iii)** statistical or data mining **analysis**:
  - Analysis may be performed by external tools (e.g. TIGR MeV, Weka, R, etc.), but this requires data movement and further data reorganization
- **(iv)** biological **interpretation** and access to **external knowledge bases**

OPEN PROBLEM: THIS PROCESS INVOLVES SEVERAL TOOLS AND CHIP-SPECIFIC / VENDOR-SPECIFIC LIBRARIES AND MAY BE ERROR PRONE

The format or raw .cel files



Affymetrix Power Tools

- The preprocessing of Affymetrix CEL files is performed using specialized tools, e.g. the APT (Affymetrix Power Tools) command line tools
- **apt-probeset-summarize** is an APT tool for summarizing and normalizing expression probe data from CEL files.
  - It needs library files (either a cdf file or pgf/clf files) for defining probesets
  - It can perform different types of summarization, e.g. the Robust Multiarray Average (RMA) and the Probe Logarithmic Intensity Error (PLIER) algorithms
    - For exon arrays, the Detection Above BackGround (DABG) is provided
  - It can perform two main types of normalization, the quantile algorithm and the sketch-quantile algorithm, that re-scale each expression value of a dataset.

```
apt-probeset-summarize -a rma-sketch -a plier-mm-sketch -d chip.cdf
-o output-dir --cel-files cel_list.txt
```

```
apt-probeset-summarize -a rma -d HuEx-1_0-st-v2.cdf -o/home/output -
cel-files/home/list.txt
```

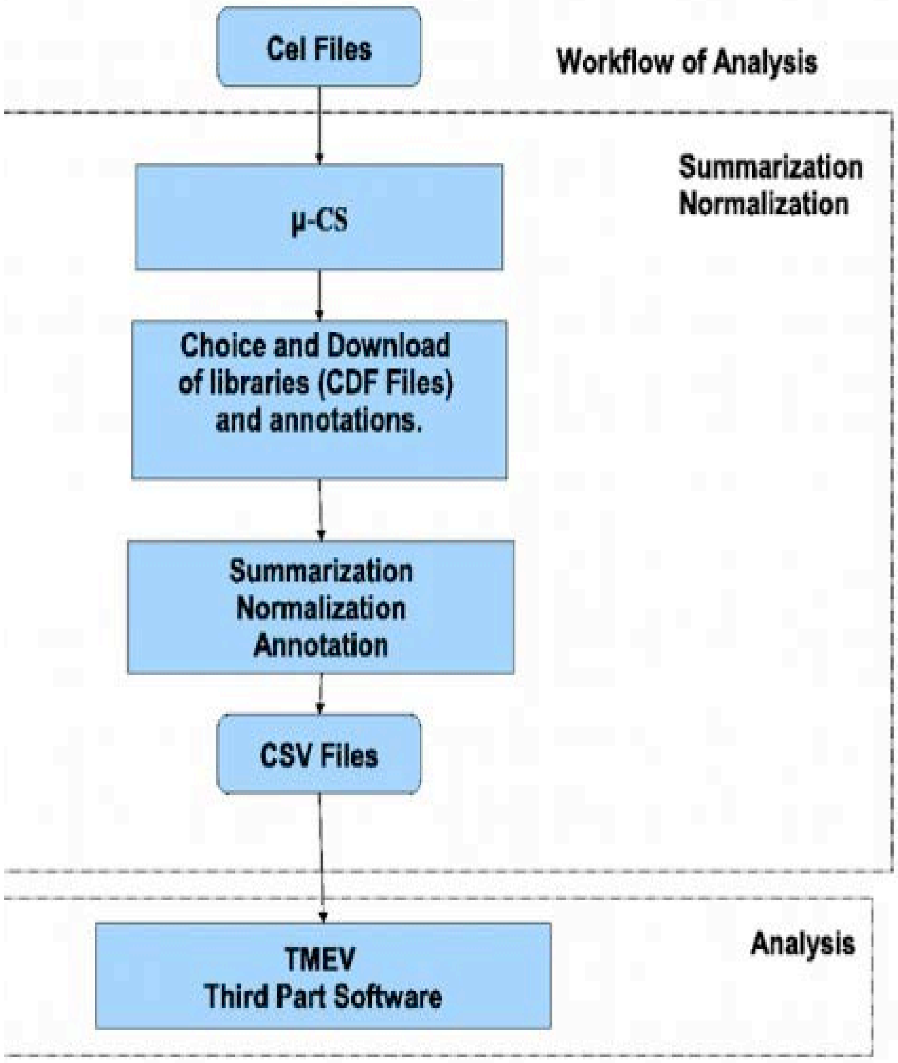
- A further step is **annotation,** that allows to associate to each expression value the related gene and further biological annotation,
  - e.g. database identifier, description of molecular function, associated protein domains, Gene Ontology data,
- by using annotation files often provided by chip manufacturer

Addressed problems

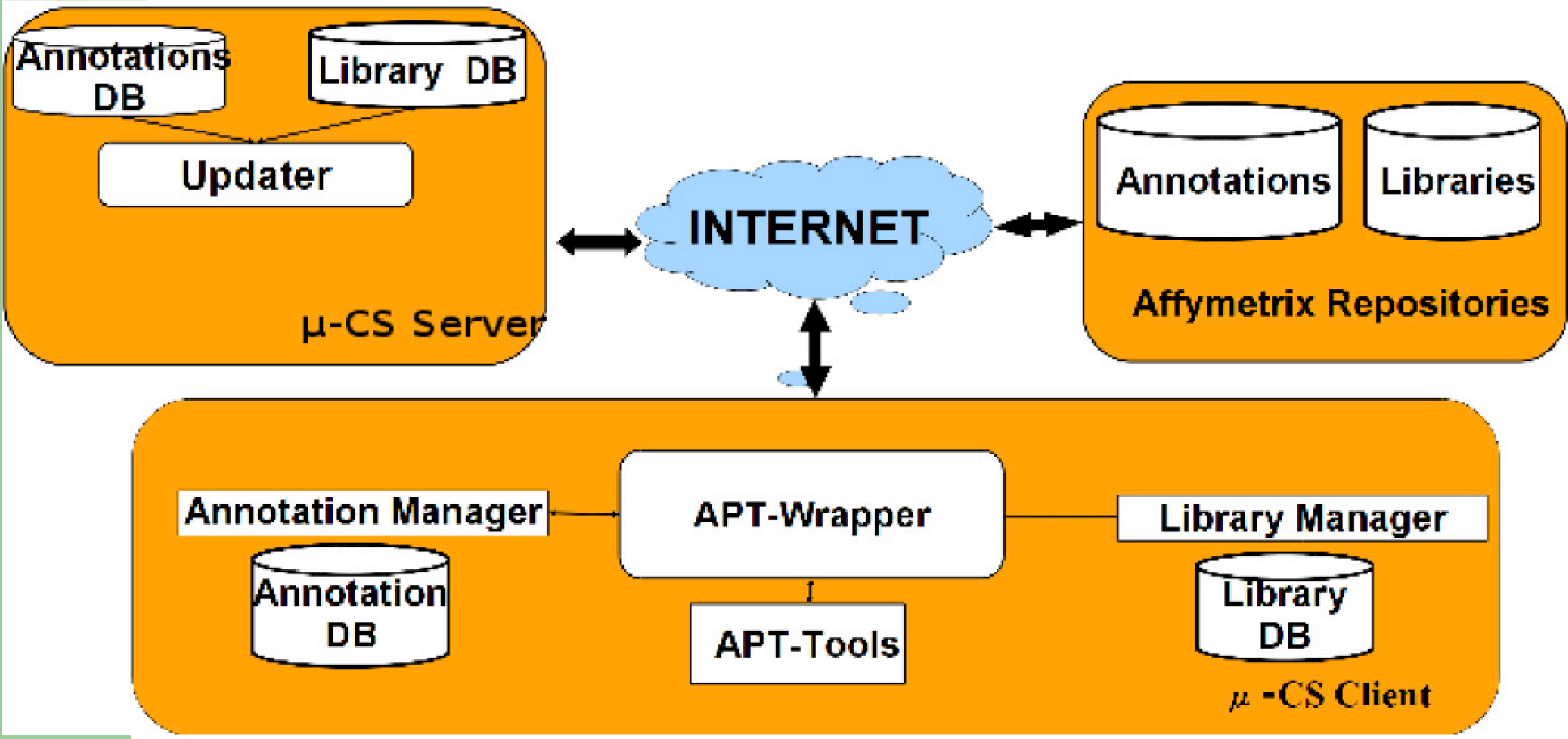
- The main drawbacks of such an approach are the need:
  - to generate and store intermediate files in a manual way that prevent the automation of the process;
  - to know details of the used chips and related preprocessing tools and libraries;
  - to manually download the most updated summarization and annotation libraries from the vendor website, and
  - to manually import preprocessed files in the analysis platforms (e.g. TM4 MeV, Weka, etc.), that may introduce errors.
- **The automation of the preprocessing pipeline could speed-up the entire analysis process and reduce possible errors, allowing the user to concentrate on biological aspects.**
- **micro-CS** is a tool to normalize, summarize and annotate gene expression data produced by Affymetrix microarray reducing user intervention
  - system handles the guided selection and automatic upgrading of the software libraries needed by Affymetrix tools to preprocess and annotate gene expressions
  - the tool is available with a stand-alone user-interface or as a TM4 MeV plug-in
    - TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003 Feb;34(2):374-8.

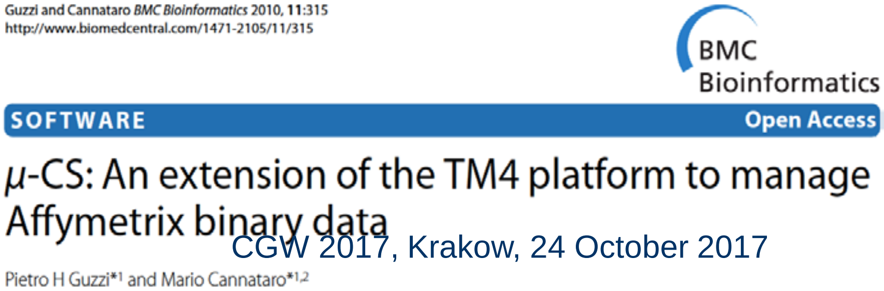
micro-CS

- μ-CS preprocesses Affymetrix microarray data to automatize summarization, normalization, and annotation of microarray.
  - μ-CS users can directly manage binary data without worrying about locating and invoking the proper preprocessing tools and chip-specific libraries.
  - users of the μ-CS plugin for TM4 can manage Affymetrix binary files without using external tools, such as APT (Affymetrix Power Tools) and related libraries.
- μ-CS offers four main advantages:
  - (i) it avoids to waste time for searching the correct libraries,
  - (ii) it reduces possible errors in the preprocessing and further analysis phases, e.g. due to the incorrect choice of parameters or the use of old libraries,
  - (iii) it implements the annotation of preprocessed data,
  - (iv) it may enhance the quality of further analysis since it provides the most updated annotation libraries.
- μ-CS client is freely available (TM4 plugin or standalone tool)



micro-CS Architecture




Guzzi and Cannataro *BMC Bioinformatics* 2010, **11**:315
http://www.biomedcentral.com/1471-2105/11/315

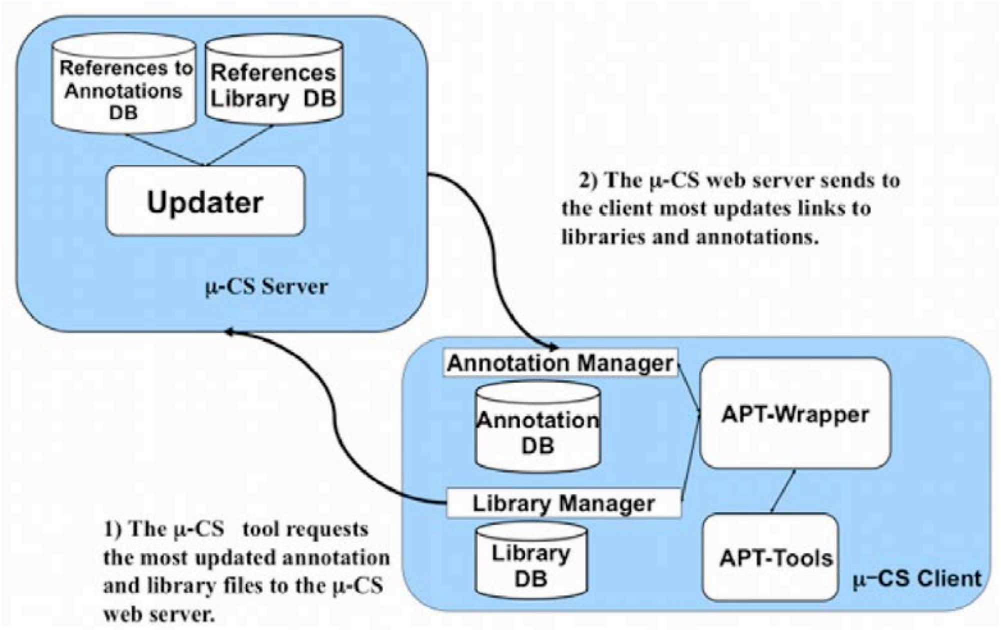
BMC Bioinformatics

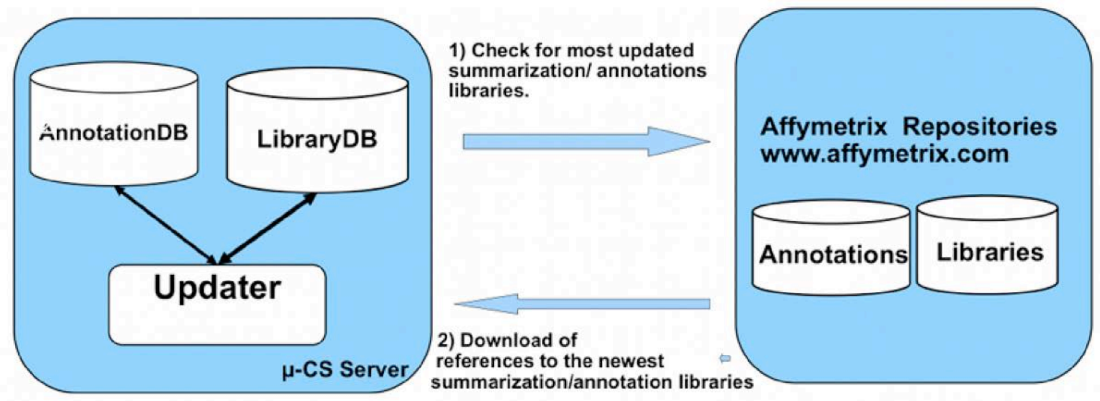
SOFTWARE **Open Access**

µ-CS: An extension of the TM4 platform to manage Affymetrix binary data

Pietro H Guzzi*[1] and Mario Cannataro*[1,2]

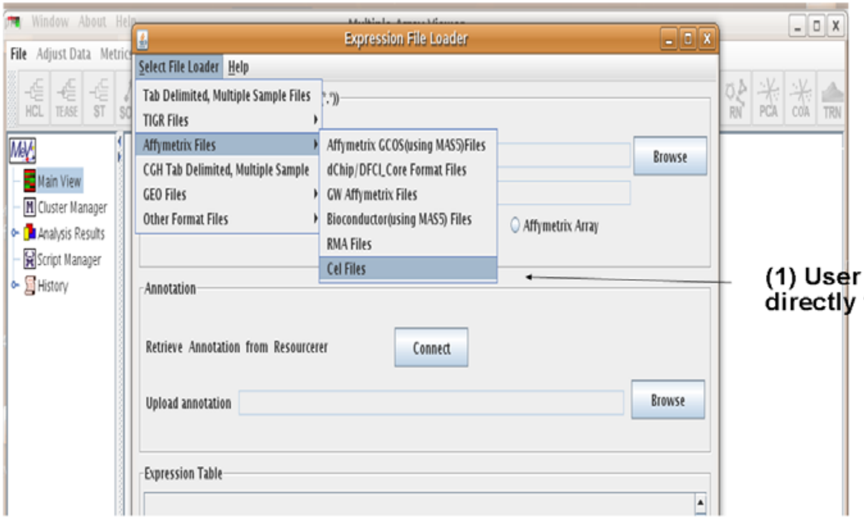
Update of the Client Databases


References to Annotations DB
References Library DB
Updater
µ-CS Server
2) The µ-CS web server sends to the client most updates links to libraries and annotations.
Annotation Manager
Annotation DB
APT-Wrapper
Library Manager
Library DB
APT-Tools
µ-CS Client
1) The µ-CS tool requests the most updated annotation and library files to the µ-CS web server.

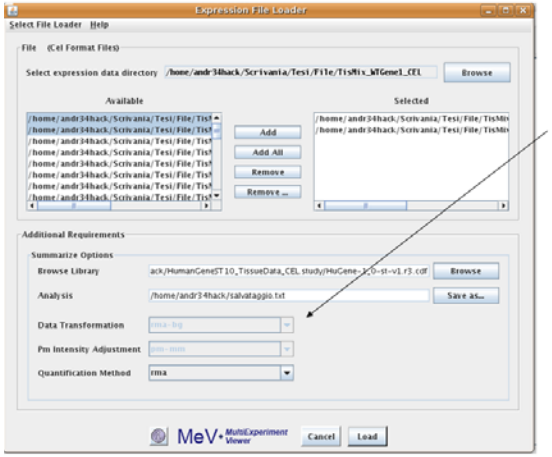

AnnotationDB
LibraryDB
Updater
µ-CS Server
1) Check for most updated summarization/ annotations libraries.
2) Download of references to the newest summarization/annotation libraries
Affymetrix Repositories www.affymetrix.com
Annotations
Libraries

Update of the Server Databases

micro-CS User Interface (TM4 plugin version)



(1) User can load transparently Cel files directly from the TMeV Loader (Madam).



File (Cel Format Files)

Select expression data directory: /home/andr34hack/Scrivania/Tesi/File/TisMix_WTGene1_CEL

Available:
/home/andr34hack/Scrivania/Tesi/File/Tis
/home/andr34hack/Scrivania/Tesi/File/Tis
/home/andr34hack/Scrivania/Tesi/File/Tis
/home/andr34hack/Scrivania/Tesi/File/Tis
/home/andr34hack/Scrivania/Tesi/File/Tis
/home/andr34hack/Scrivania/Tesi/File/Tis
/home/andr34hack/Scrivania/Tesi/File/Tis
/home/andr34hack/Scrivania/Tesi/File/Tis

Selected:
/home/andr34hack/Scrivania/Tesi/File/TisMix
/home/andr34hack/Scrivania/Tesi/File/TisMix

Additional Requirements

Summarize Options

Browse Library: ack/HumanGeneST10_TissueData_CEL study/HuGene-1_0-st-v1.r3.cdf

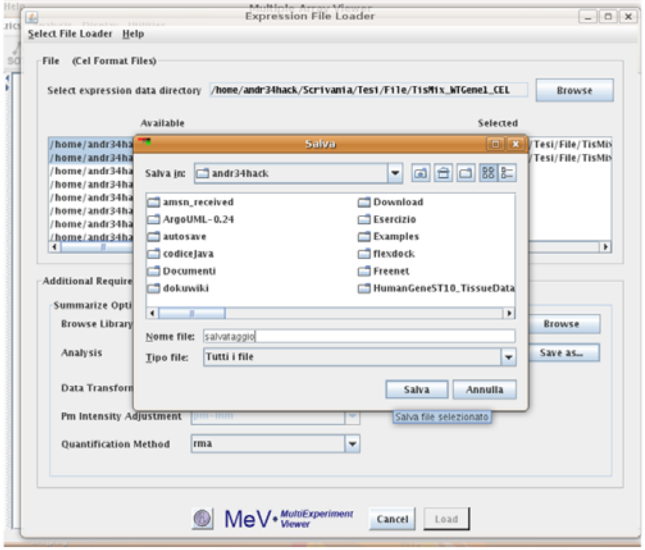
Analysis: /home/andr34hack/salvataggio.txt

Data Transformation: rma-bg

Pm Intensity Adjustment: pm-mm

Quantification Method: rma

(2) Then he/she has to set main parameters of summarization (e.g. kind of summarization and related parameters)



File (Cel Format Files)

Select expression data directory: /home/andr34hack/Scrivania/Tesi/File/TisMix_WTGene1_CEL

Salva in: andr34hack

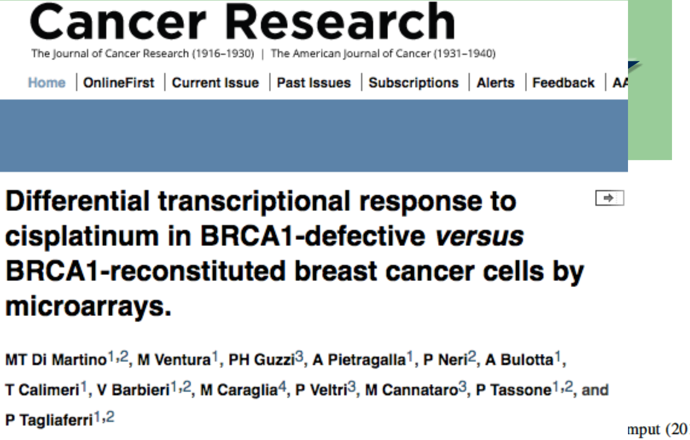
amsn_received
ArgoUML-0.24
autosave
codiceJava
Documenti
dokuwiki
Download
Esercizio
Examples
flexdock
Freenet
HumanGeneST10_TissueData

Nome file: salvataggio

Tipo file: Tutti i file

Quantification Method: rma

(3) Finally, processed files are ready to be exported in a file or loaded in TMeV



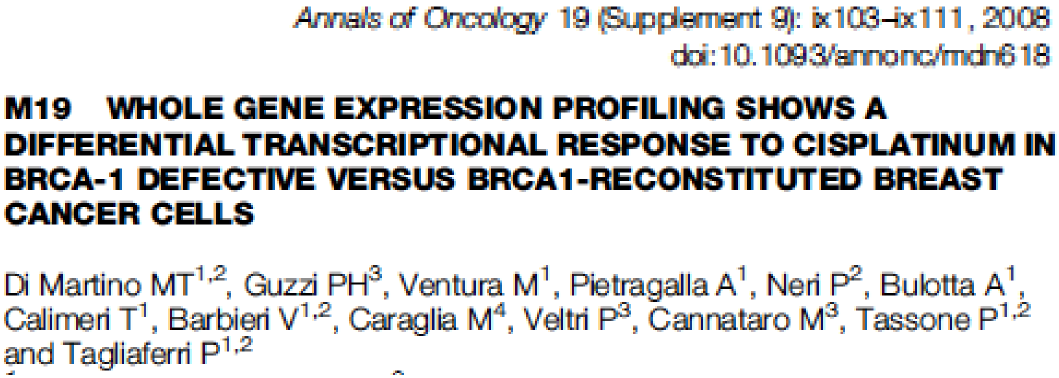
Cancer Research

The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

Home | OnlineFirst | Current Issue | Past Issues | Subscriptions | Alerts | Feedback | AA

Differential transcriptional response to cisplatinum in BRCA1-defective *versus* BRCA1-reconstituted breast cancer cells by microarrays.

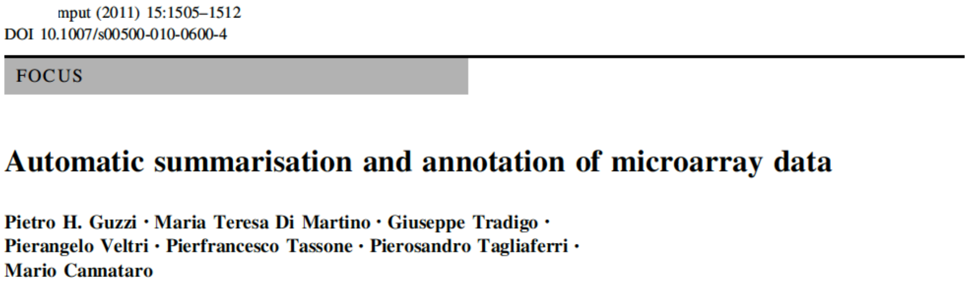
MT Di Martino[1,2], M Ventura[1], PH Guzzi[3], A Pietragalla[1], P Neri[2], A Bulotta[1], T Calimeri[1], V Barbieri[1,2], M Caraglia[4], P Veltri[3], M Cannataro[3], P Tassone[1,2], and P Tagliaferri[1,2]



Annals of Oncology 19 (Supplement 9): ix103–ix111, 2008
doi:10.1093/annonc/mdn618

M19 WHOLE GENE EXPRESSION PROFILING SHOWS A DIFFERENTIAL TRANSCRIPTIONAL RESPONSE TO CISPLATINUM IN BRCA-1 DEFECTIVE VERSUS BRCA1-RECONSTITUTED BREAST CANCER CELLS

Di Martino MT[1,2], Guzzi PH[3], Ventura M[1], Pietragalla A[1], Neri P[2], Bulotta A[1], Calimeri T[1], Barbieri V[1,2], Caraglia M[4], Veltri P[3], Cannataro M[3], Tassone P[1,2] and Tagliaferri P[1,2]



mput (2011) 15:1505–1512
DOI 10.1007/s00500-010-0600-4

FOCUS

Automatic summarisation and annotation of microarray data

Pietro H. Guzzi · Maria Teresa Di Martino · Giuseppe Tradigo · Pierangelo Veltri · Pierfrancesco Tassone · Pierosandro Tagliaferri · Mario Cannataro

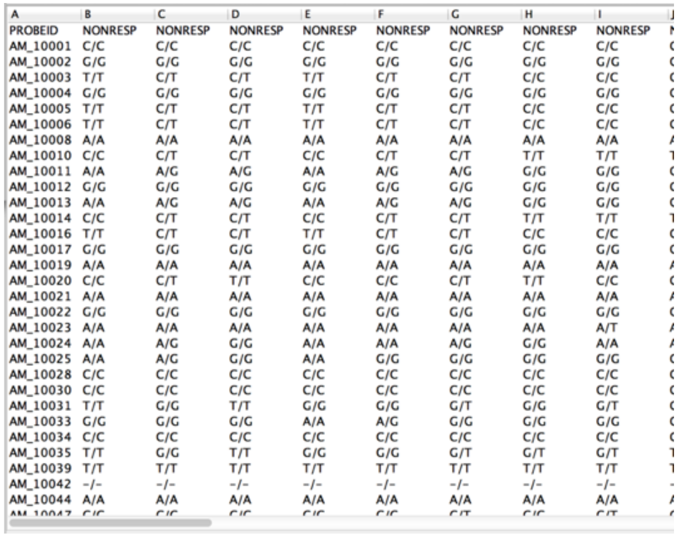
- To study the molecular bases of BRCA1-related differential sensitivity to the drug, we analyzed the whole gene expression profile of HCC1937 and HCC1937/wtBRCA1 cells following in vivo and in vitro exposure of tumor cells to cisplatinum (CDDP)
  - HCC1937 is a BRCA1-defective breast cancer cell line which discloses higher sensitivity to CDDP as compared to the BRCA1 full length cDNA transfected clone HCC1937/wtBRCA1
  - Gene expression profiling was performed by Affymetrix technology using **Human GeneArray 1.0ST**.
  - **Array data were preprocessed using µ-CS** and analyzed using Gene Expression Console, GeneSpring and Ingenuity Pathway Analysis (IPA).

Part I: Genomics Data

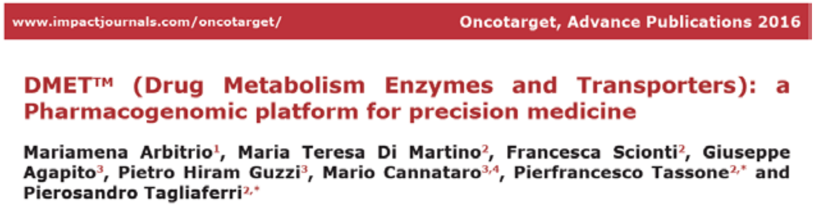
- Microarray-based Genomics Data
- Automatic preprocessing of gene expression data
  - **micro-CS**: preprocessing and annotation of gene expression data
- Statistical analysis of genotyping data
  - **DMET-Analyzer**: preprocessing and analysis of DMET (Drug Metabolizing Enzymes and Transporters) SNP (Single Nucleotide Polymorphism) data
  - **coreSNP** (multicore implementation)
- Data Mining analysis of genotyping data
  - **DMET-Miner**: Association Rule Mining to extract from DMET data Association Rules able to correlate the contemporary presence of SNPs with patient's conditions (e.g. TOX vs NOTOX)

Affymetrix DMET Plus Premier Pack Microarray

- The **Affymetrix DMET** (Drug Metabolism Enzymes and Transporters) platform **investigates 225 ADME genes**, i.e. genes involved in Absorption, Distribution, Metabolism and Excretion (ADME) of drugs, **for pharmacogenomics case-control study**.
- **Pharmacogenomics** is a branch of genomics that aims to predict the response to drugs of an individual based on an his/her genotype
  - **The hypothesis of DMET analysis is that a different response to drugs may be related to modifications (SNPs) in those ADME genes**
- We want identify the SNPs that make effective/ineffective a drug treatment using efficient Statistical Analysis and Association Rule Mining



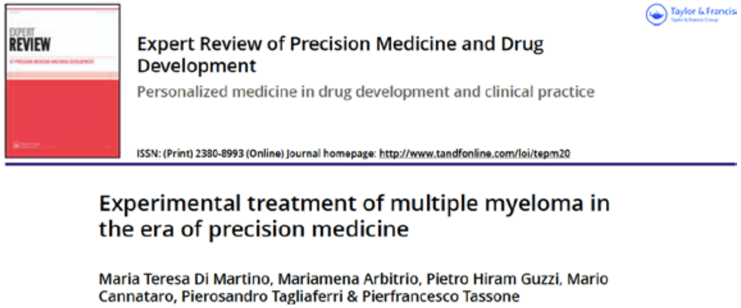
A	B	C	D	E	F	G	H	I
PROBEID	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP
AM_10001	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10002	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10003	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C
AM_10004	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10005	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C
AM_10006	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C
AM_10008	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10010	C/C	C/T	C/T	C/C	C/T	C/T	T/T	T/T
AM_10011	A/A	A/G	A/G	A/A	A/G	A/G	G/G	G/G
AM_10012	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10013	A/A	A/G	A/G	A/A	A/G	A/G	G/G	G/G
AM_10014	C/C	C/T	C/T	C/C	C/T	C/T	T/T	T/T
AM_10016	T/T	C/T	C/T	T/T	C/T	C/T	C/C	C/C
AM_10017	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10019	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10020	C/C	C/T	T/T	C/C	C/C	C/T	T/T	C/C
AM_10021	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10022	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
AM_10023	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/T
AM_10024	A/A	A/G	G/G	A/A	A/A	A/G	G/G	A/A
AM_10025	A/A	A/G	G/G	A/A	G/G	G/G	G/G	G/G
AM_10028	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10030	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10031	T/T	G/G	T/T	G/G	G/G	G/T	G/G	G/T
AM_10033	G/G	G/G	G/G	A/A	A/G	G/G	G/G	G/G
AM_10034	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
AM_10035	T/T	G/G	T/T	G/G	G/G	G/T	G/T	G/T
AM_10039	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
AM_10042	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
AM_10044	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
AM_10047	C/C	C/C	C/C	C/C	C/C	C/T	C/C	C/T



www.impactjournals.com/oncotarget/ Oncotarget, Advance Publications 2016

DMET™ (Drug Metabolism Enzymes and Transporters): a Pharmacogenomic platform for precision medicine

Mariamena Arbitrio[1], Maria Teresa Di Martino[2], Francesca Scionti[2], Giuseppe Agapito[3], Pietro Hiram Guzzi[3], Mario Cannataro[3,4], Pierfrancesco Tassone[2,*] and Pierosandro Tagliaferri[2,*]



EXPERT REVIEW

Taylor & Francis

Expert Review of Precision Medicine and Drug Development

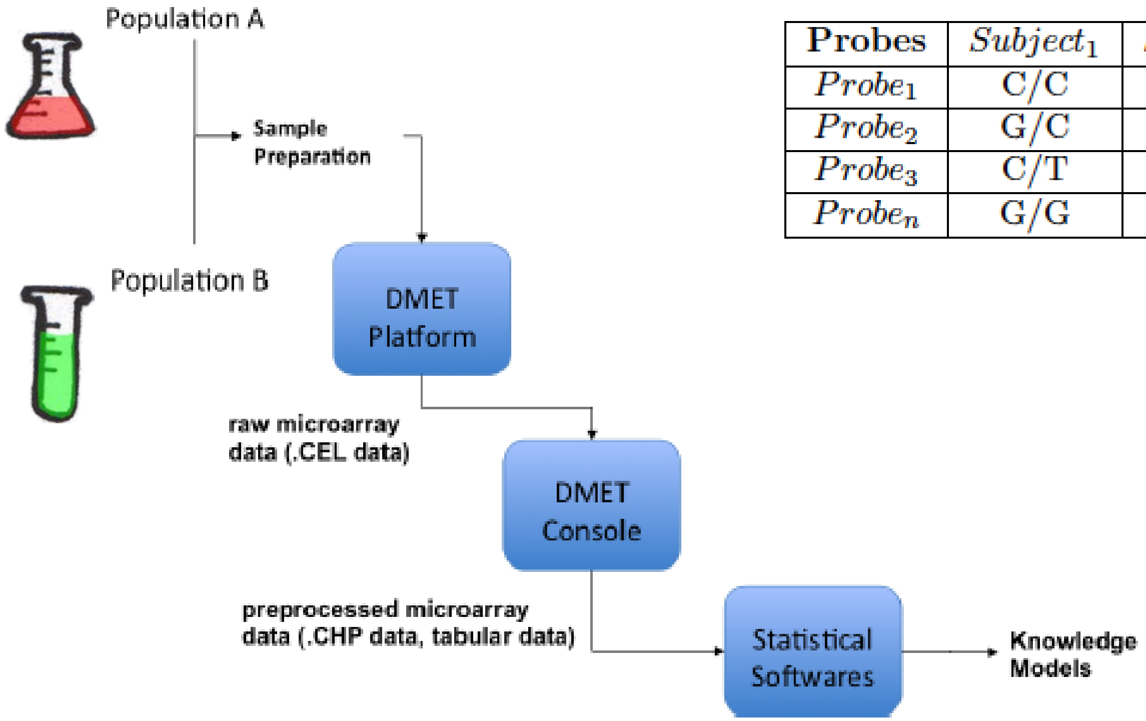
Personalized medicine in drug development and clinical practice

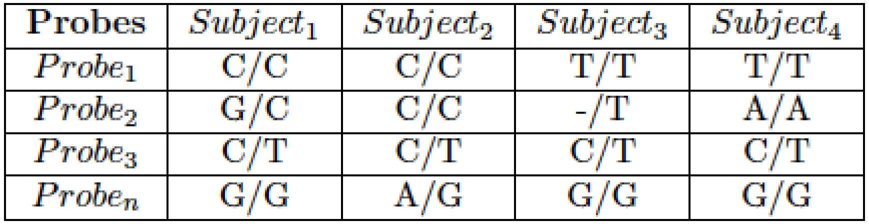
ISSN: (Print) 2380-8993 (Online) Journal homepage: http://www.tandfonline.com/loi/tepm20

Experimental treatment of multiple myeloma in the era of precision medicine

Maria Teresa Di Martino, Mariamena Arbitrio, Pietro Hiram Guzzi, Mario Cannataro, Pierosandro Tagliaferri & Pierfrancesco Tassone

Workflow of a DMET pharmacogenomics experiment





Probes	$Subject_1$	$Subject_2$	$Subject_3$	$Subject_4$
$Probe_1$	C/C	C/C	T/T	T/T
$Probe_2$	G/C	C/C	-/T	A/A
$Probe_3$	C/T	C/T	C/T	C/T
$Probe_n$	G/G	A/G	G/G	G/G

- **OPEN PROBLEMS:**
  - **NO BIOINFORMATICS SOFTWARE FOR ANALYZING DMET SNP DATA WAS AVAILABLE AT THE TIME OF RESEARCH (ONLY MANUAL ANALYSIS WITH EXTERNAL SOFTWARE)**
  - **PERFORMANCE ISSUES WHEN INCREASING SIZE OF EXAMINED POPULATION**

Workflow of a DMET pharmacogenomics experiment

- **Sample collection and DMET chip preparation**: biological samples are collected and treated to perform microarray experiments;
  - Affymetrix DMET chip allows the investigation of 1936 probes, each one representing a portion of the genome having a role in drug metabolism;
  - usually samples are divided in two classes, e.g. on the basis of the response to a drug (Population A and B).
- **DMET microarray experiments**: produce raw expression data, contained in .CEL files (one file per sample);
- **DMET data preprocessing and SNPs detection**: Affymetrix proprietary tools, (e.g. **apt-dmet-genotype** command line tool or **DMET Console**) are used to summarize and average expression values contained in .CEL file to produce .CHP files (one file per sample);
  - all .CHP files need to be combined to form a single table containing the detected SNPs for all samples;
  - similarly to gene expression microarray, the SNPs table will contain in position (i,j) the SNP detected by probe i in the sample j;
- **SNPs analysis**: usually the Fisher statistical test is used to evaluate if a different distribution of SNPs among two classes of samples is statistically significant or not. The analysis, often performed manually using external software, produces a list of SNPs ordered by p-value

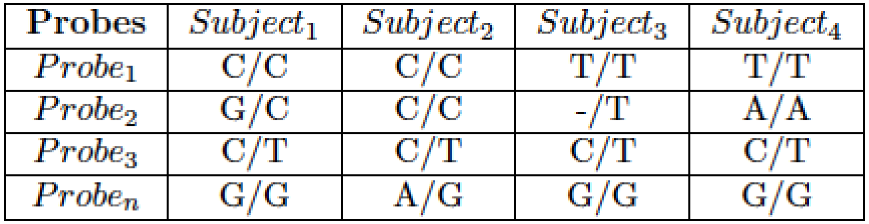
DMET-Analyzer

- DMET-Analyzer is a tool for automated pre-processing and statistical analysis of SNP data generated by the Affymetrix DMET platform
  - the system finds and highlights all statistically significant SNP variations found in the data
- DMET-Analyzer supports the automatic statistical analysis in DMET-based pharmacogenomics studies.
  - It has a simple graphical user interface that allows users (doctors/biologists) to upload and analyze DMET files produced by DMET Console in an interactive way.
- Starting from a DMET dataset, DMET-Analyzer is able to find statistically relevant subsets of SNPs that separate two input classes
  - DMET-Analyzer tests in an automatic way if the distributions of each SNPs detected on each probe on the two classes of subjects (e.g. Healthy vs Diseased, RESP vs NORESP, TOX vs NOTOX) are statistically significant.
- The system is freely available and currently used by the Oncology Unit of the “Mater Domini” University Hospital, Catanzaro, Italy.

Algorithm

DMET-Analyzer allows to test in an automatic way if the different distributions of each SNPs detected on each probe on the two classes of subjects (e.g. RESP vs NORESP) are statistically significant.

- It performs a massive amount of Fisher’s tests without user intervention
- It provides several statistical corrections (FDR, Bonferroni) useful for low samples experiments
- It uses two parameters to
  - discard probes whose SNP distributions are “similar”, avoiding useless tests
  - discard Fisher’s tests with high p-value



Probes	$Subject_1$	$Subject_2$	$Subject_3$	$Subject_4$
$Probe_1$	C/C	C/C	T/T	T/T
$Probe_2$	G/C	C/C	-/T	A/A
$Probe_3$	C/T	C/T	C/T	C/T
$Probe_n$	G/G	A/G	G/G	G/G

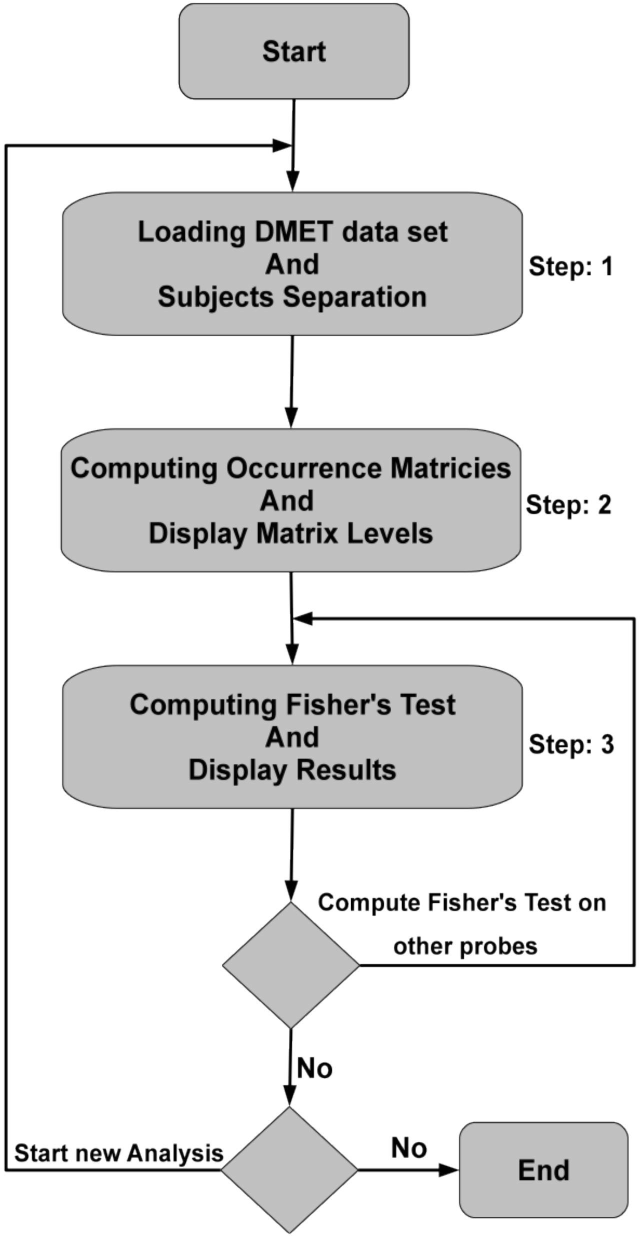
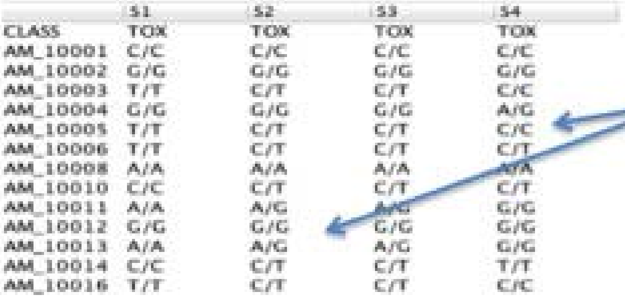


Fig. 2. DMET-Analyser flowchart.

Data Analysis Workflow

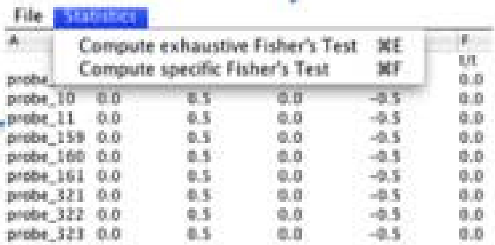
a User chooses the Data File to be loaded.



	S1	S2	S3	S4
CLASS	TOX	TOX	TOX	TOX
AM_10001	C/C	C/C	C/C	C/C
AM_10002	G/G	G/G	G/G	G/G
AM_10003	T/T	C/T	C/T	C/C
AM_10004	G/G	G/G	G/G	A/G
AM_10005	T/T	C/T	C/T	C/C
AM_10006	T/T	C/T	C/T	C/T
AM_10008	A/A	A/A	A/A	A/A
AM_10010	C/C	C/T	C/T	C/T
AM_10011	A/A	A/G	A/G	G/G
AM_10012	G/G	G/G	G/G	G/G
AM_10013	A/A	A/G	A/G	G/G
AM_10014	C/C	C/T	C/T	T/T
AM_10016	T/T	C/T	C/T	C/C

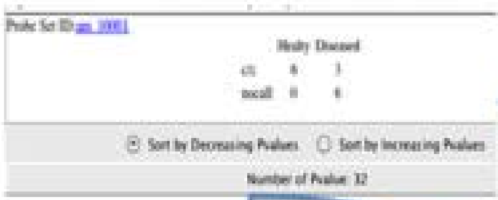
b User selects the classes for each sample by clicking on data table.

c Allele frequencies for each probe are calculated and shown to the user


File
Statistics
Compute exhaustive Fisher's Test ⌘E
Compute specific Fisher's Test ⌘F

A					F
					0.0
probe_[illegible]					0.0
probe_10	0.0	0.5	0.0	-0.5	0.0
probe_11	0.0	0.5	0.0	-0.5	0.0
probe_159	0.0	0.5	0.0	-0.5	0.0
probe_160	0.0	0.5	0.0	-0.5	0.0
probe_161	0.0	0.5	0.0	-0.5	0.0
probe_321	0.0	0.5	0.0	-0.5	0.0
probe_322	0.0	0.5	0.0	-0.5	0.0
probe_323	0.0	0.5	0.0	-0.5	0.0

Results are calculated



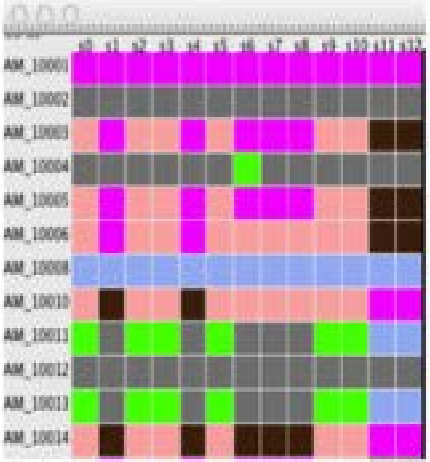
Probe Set ID: am_10001

	Healthy	Diseased
c/c	6	1
nocall	0	6

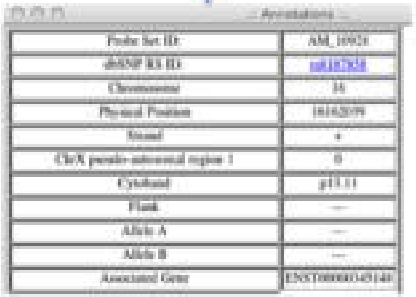
◉ Sort by Decreasing Pvalues ○ Sort by Increasing Pvalues

Number of Pvalue: 32

d User selects the Statistical Test and may visualize distribution.



e Results are shown to the user. He may visualize annotations and links to esternal databases.



Probe Set ID	AM_10021
dbSNP RS ID	rs[illegible]
Chromosome	16
Physical Position	16162079
Strand	+
ChrX pseudo-autosomal region 1	0
Cytoband	p13.11
Flank	---
Allele A	---
Allele B	---
Associated Gene	ENST00000345146

Dubium s

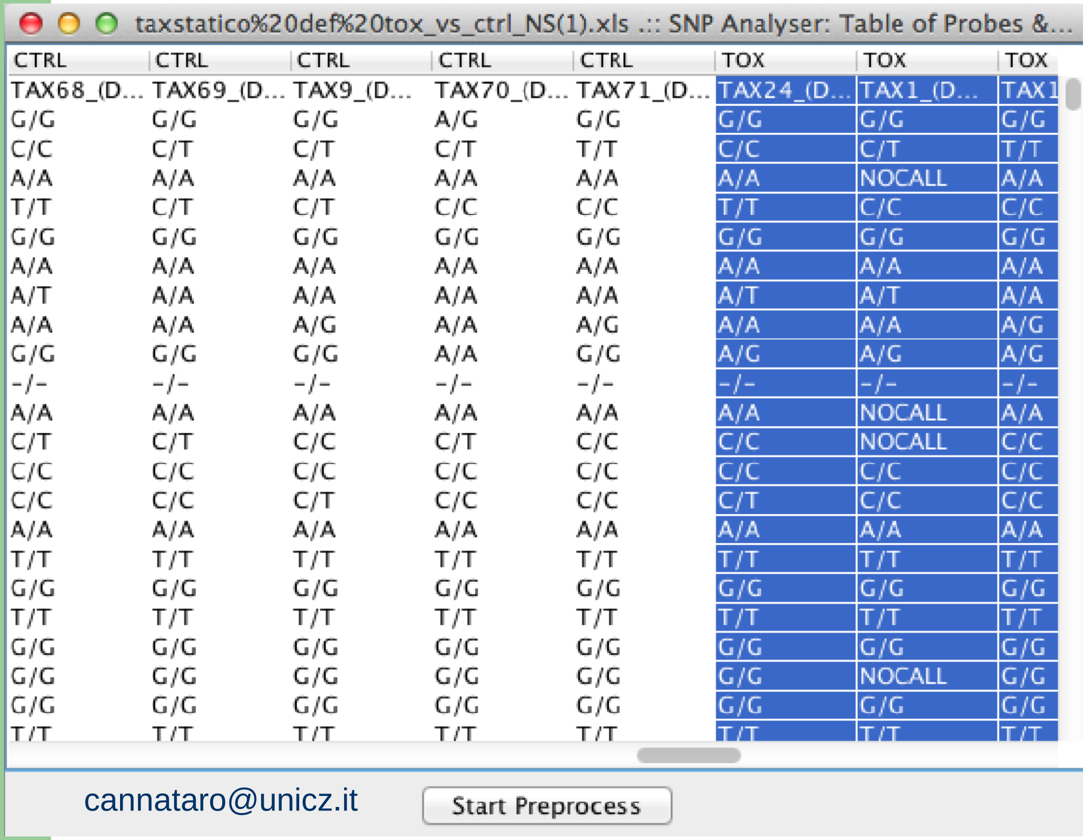
SOFTWARE Open Access

DMET-Analyzer: automatic analysis of Affymetrix DMET Data

BMC Bioinformatics

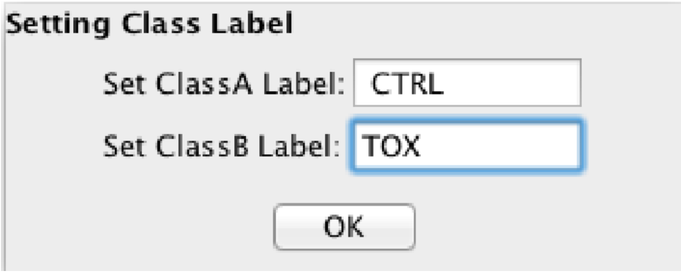
Pietro Hiram Guzzi[1*], Giuseppe Agapito[1], Maria Teresa Di Martino[2], Mariamena Arbitrio[3], Pierfrancesco Tassone[2], Pierosandro Tagliaferri[2] and Mario Cannataro[1]

DMET-Analyzer use: loading input data and selection of classes



taxstatico%20def%20tox_vs_ctrl_NS(1).xls .:: SNP Analyser: Table of Probes &...

CTRL	CTRL	CTRL	CTRL	CTRL	TOX	TOX	TOX
TAX68_(D...	TAX69_(D...	TAX9_(D...	TAX70_(D...	TAX71_(D...	TAX24_(D...	TAX1_(D...	TAX1
G/G	G/G	G/G	A/G	G/G	G/G	G/G	G/G
C/C	C/T	C/T	C/T	T/T	C/C	C/T	T/T
A/A	A/A	A/A	A/A	A/A	A/A	NOCALL	A/A
T/T	C/T	C/T	C/C	C/C	T/T	C/C	C/C
G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
A/T	A/A	A/A	A/A	A/A	A/T	A/T	A/A
A/A	A/A	A/G	A/A	A/G	A/A	A/A	A/G
G/G	G/G	G/G	A/A	G/G	A/G	A/G	A/G
-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
A/A	A/A	A/A	A/A	A/A	A/A	NOCALL	A/A
C/T	C/T	C/C	C/T	C/C	C/C	NOCALL	C/C
C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
C/C	C/C	C/T	C/C	C/C	C/T	C/C	C/C
A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
G/G	G/G	G/G	G/G	G/G	G/G	NOCALL	G/G
G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T

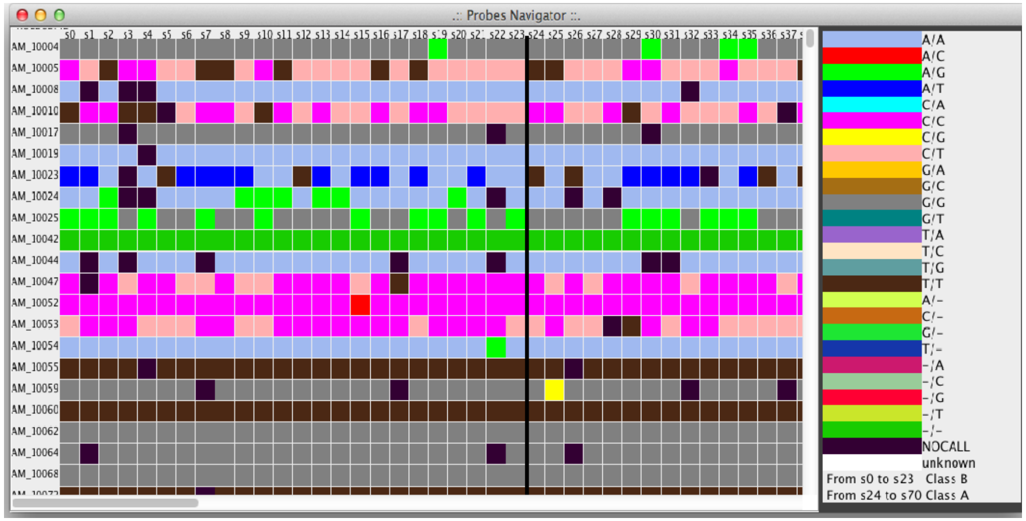


Setting Class Label

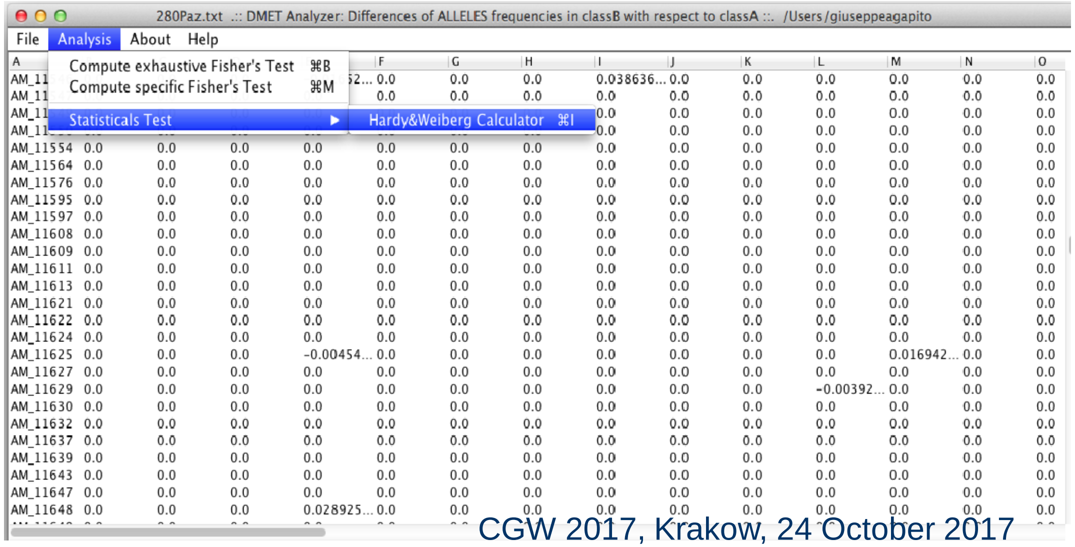
Set ClassA Label: CTRL

Set ClassB Label: TOX

Heat Map probe navigator window



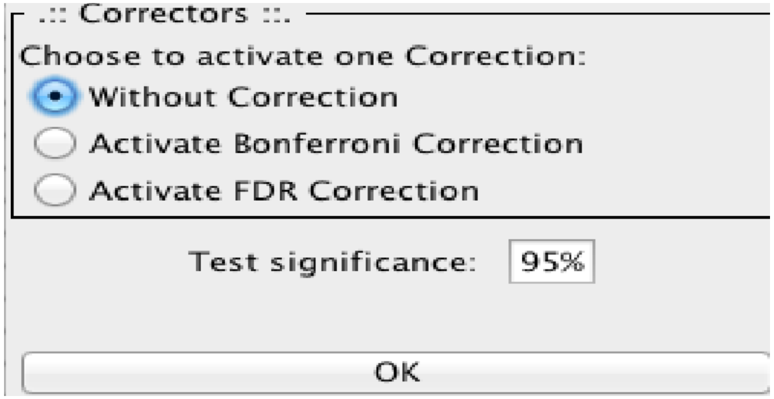
Frequecy Differences Table - differences of SNPs frequencies for each probe (class B w.r.t class A)


File
Analysis
About
Help
Compute exhaustive Fisher's Test ⌘B
Compute specific Fisher's Test ⌘M
Statisticals Test ▶
Hardy&Weiberg Calculator ⌘I

280Paz.txt .:: DMET Analyzer: Differences of ALLELES frequencies in classB with respect to classA ::. /Users/giuseppeagapito

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
AM_11546	0.0	0.0	0.0	-0.01652...	0.0	0.0	0.0	0.038636...	0.0	0.0	0.0	0.0	0.0	0.0
AM_11547	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11554	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11564	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11576	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11595	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11597	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11608	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11609	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11611	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11613	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11621	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11622	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11624	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11625	0.0	0.0	0.0	-0.00454...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.016942...	0.0	0.0
AM_11627	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11629	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.00392...	0.0	0.0	0.0
AM_11630	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11632	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11637	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11639	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11643	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11647	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AM_11648	0.0	0.0	0.0	0.028925...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Statistical tests:



.:: Correctors ::.

Choose to activate one Correction:

● Without Correction

○ Activate Bonferroni Correction

○ Activate FDR Correction

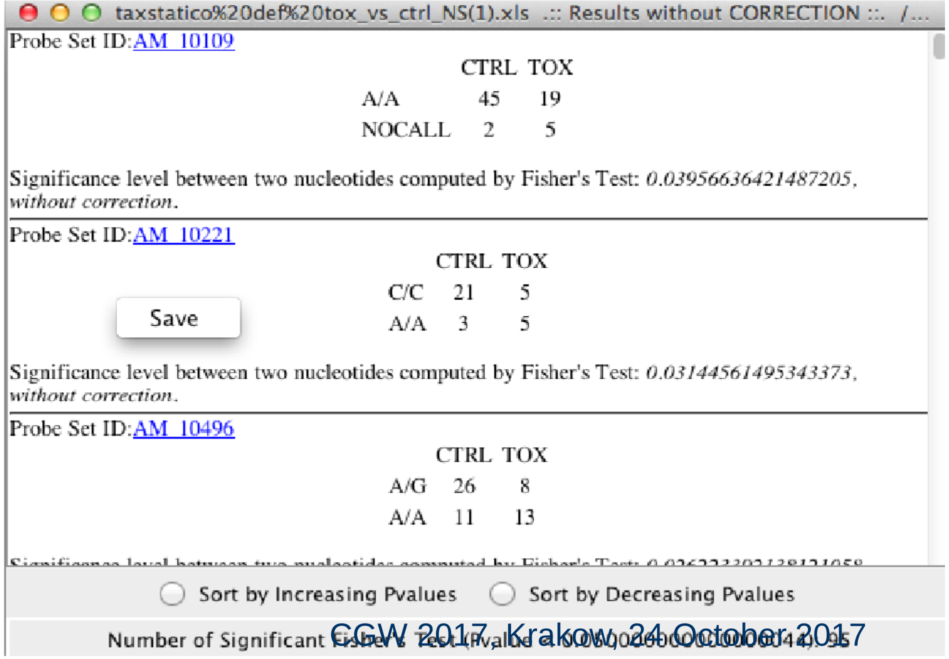
Test significance: 95%

1) Setting correction;

2) Executing Fisher’s tests;

3) Analyze results.

- The Fishers test calculator uses the data contained into two Occurrences Tables to compute the Fishers test for each couple of allele belonging to the two classes.
- The algorithm avoids the computation of trivial tests, e.g. tests where the Fisher Test contingency table contains three zero values that leads to a p_value=1 are discarded.
- Fisher tests results with *p_value > Ft* (**Filter Threshold**, default 0.05) are discarded.



taxstatico%20def%20tox_vs_ctrl_NS(1).xls .:: Results without CORRECTION ::. /...

Probe Set ID:AM_10109

	CTRL	TOX
A/A	45	19
NOCALL	2	5

Significance level between two nucleotides computed by Fisher's Test: *0.039566364214872O5, without correction.*

Probe Set ID:AM_10221

	CTRL	TOX
C/C	21	5
A/A	3	5

Significance level between two nucleotides computed by Fisher's Test: *0.031445614953433 73, without correction.*

Probe Set ID:AM_10496

	CTRL	TOX
A/G	26	8
A/A	11	13

Significance level between two nucleotides computed by Fisher's Test: 0.026223392138121058

Number of Significant Fisher's Test (Pvalue < 0.05000000000000004): 957

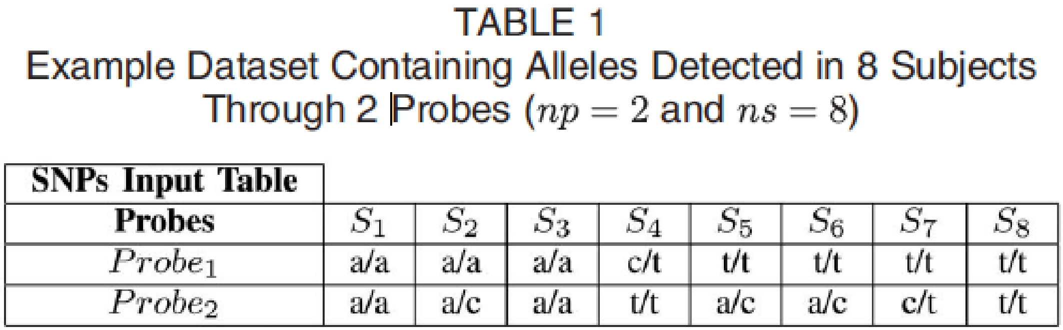


TABLE 1
Example Dataset Containing Alleles Detected in 8 Subjects Through 2 Probes ($np = 2$ and $ns = 8$)

SNPs Input Table								
Probes	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
$Probe_1$	a/a	a/a	a/a	c/t	t/t	t/t	t/t	t/t
$Probe_2$	a/a	a/c	a/a	t/t	a/c	a/c	c/t	t/t

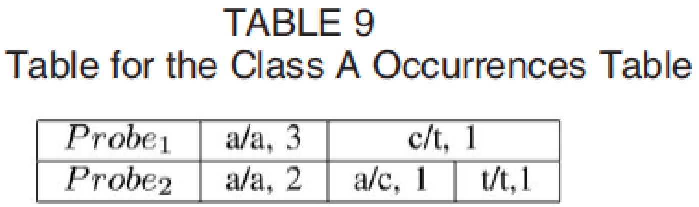


TABLE 9
Table for the Class A Occurrences Table

$Probe_1$	a/a, 3	c/t, 1	
$Probe_2$	a/a, 2	a/c, 1	t/t,1

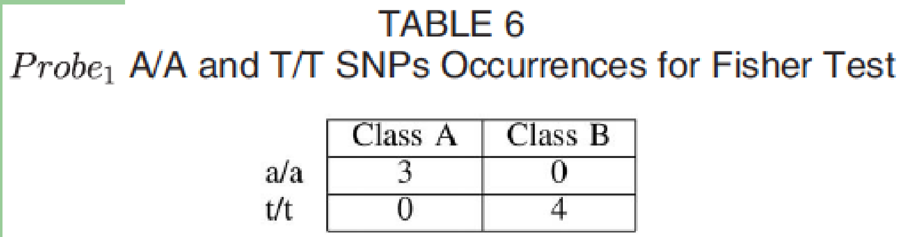


TABLE 6
$Probe_1$ A/A and T/T SNPs Occurrences for Fisher Test

	Class A	Class B
a/a	3	0
t/t	0	4

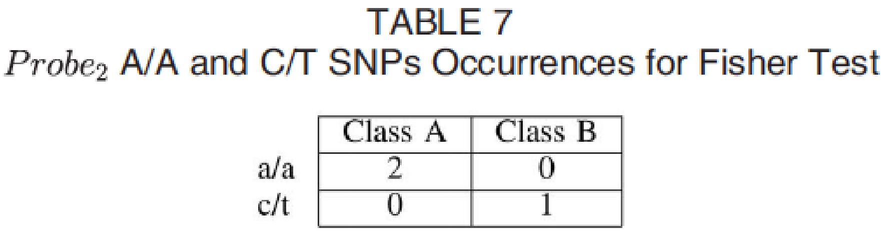


TABLE 7
$Probe_2$ A/A and C/T SNPs Occurrences for Fisher Test

	Class A	Class B
a/a	2	0
c/t	0	1

P-value = 0,0286

P-value = 0,3333

The formula below gives the exact hypergeometric probability of observing this particular arrangement of data, assuming the given marginal totals, on the null hypothesis that Class A and Class B subjects are equally likely to have those SNPs

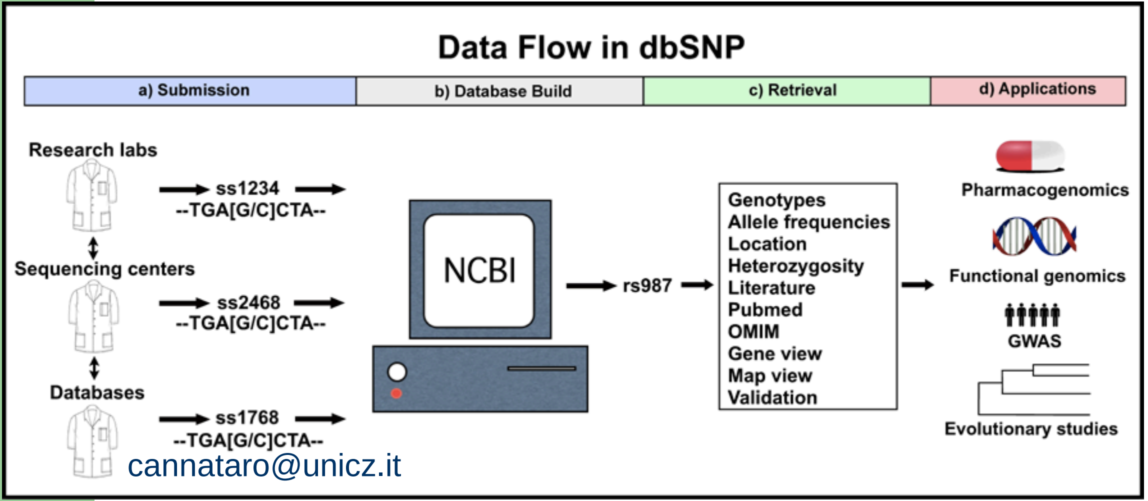
Contingency Table	Class A	Class B	
SNP 1	a	b	a+b
SNP 2	c	d	c+d
	a+c	b+d	a+b+c+d (=n)

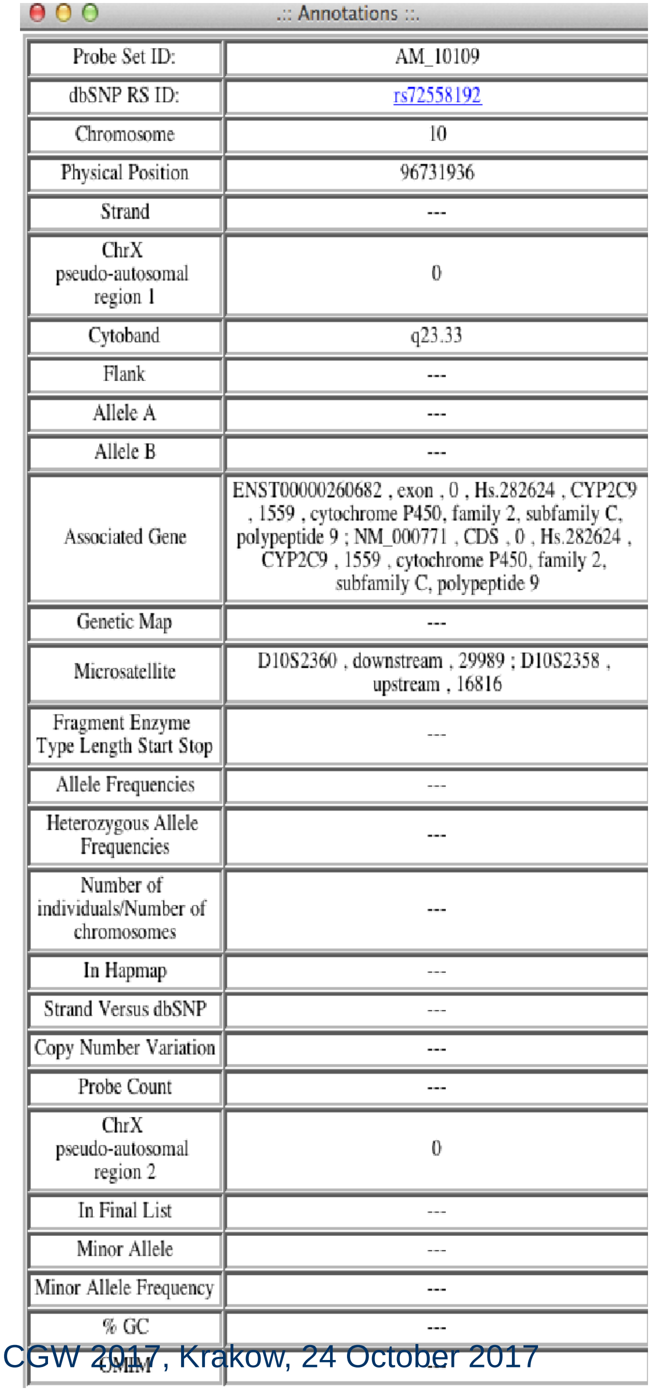
Fisher : the probability of obtaining any such set of values is given by the hypergeometric distribution

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\;(c+d)!\;(a+c)!\;(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

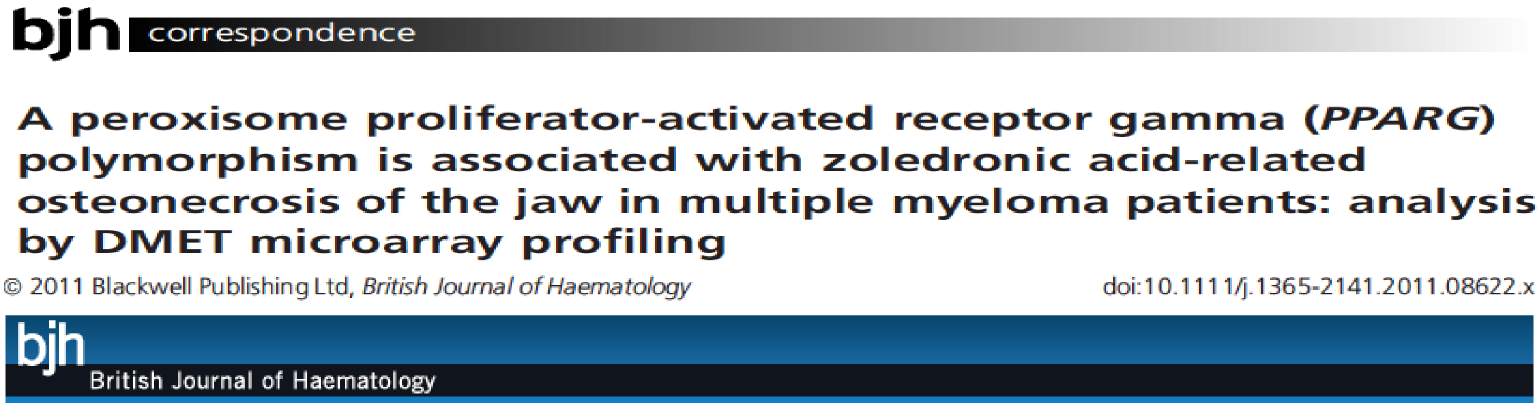
Analysis of results: annotations, dbSNP and PharmGKB links

- Each probe presented in the result window:
  - has a link that allows users to visualize annotations
  - is enriched of links to external databases such as **dbSNP** and **PharmGKB** allowing to automatically retrieve further SNP information
- PharmGKB stores knowledge about the impact of genetic variation on drug response for clinicians and researchers
- dbSNP is a public-domain archive for simple genetic polymorphisms





Probe Set ID:	AM_10109
dbSNP RS ID:	rs72558192
Chromosome	10
Physical Position	96731936
Strand	---
ChrX pseudo-autosomal region 1	0
Cytoband	q23.33
Flank	---
Allele A	---
Allele B	---
Associated Gene	ENST00000260682 , exon , 0 , Hs.282624 , CYP2C9 , 1559 , cytochrome P450, family 2, subfamily C, polypeptide 9 ; NM_000771 , CDS , 0 , Hs.282624 , CYP2C9 , 1559 , cytochrome P450, family 2, subfamily C, polypeptide 9
Genetic Map	---
Microsatellite	D10S2360 , downstream , 29989 ; D10S2358 , upstream , 16816
Fragment Enzyme Type Length Start Stop	---
Allele Frequencies	---
Heterozygous Allele Frequencies	---
Number of individuals/Number of chromosomes	---
In Hapmap	---
Strand Versus dbSNP	---
Copy Number Variation	---
Probe Count	---
ChrX pseudo-autosomal region 2	0
In Final List	---
Minor Allele	---
Minor Allele Frequency	---
% GC	---
OMIM	---



bjh correspondence

A peroxisome proliferator-activated receptor gamma (*PPARG*) polymorphism is associated with zoledronic acid-related osteonecrosis of the jaw in multiple myeloma patients: analysis by DMET microarray profiling

© 2011 Blackwell Publishing Ltd, *British Journal of Haematology* doi:10.1111/j.1365-2141.2011.08622.x

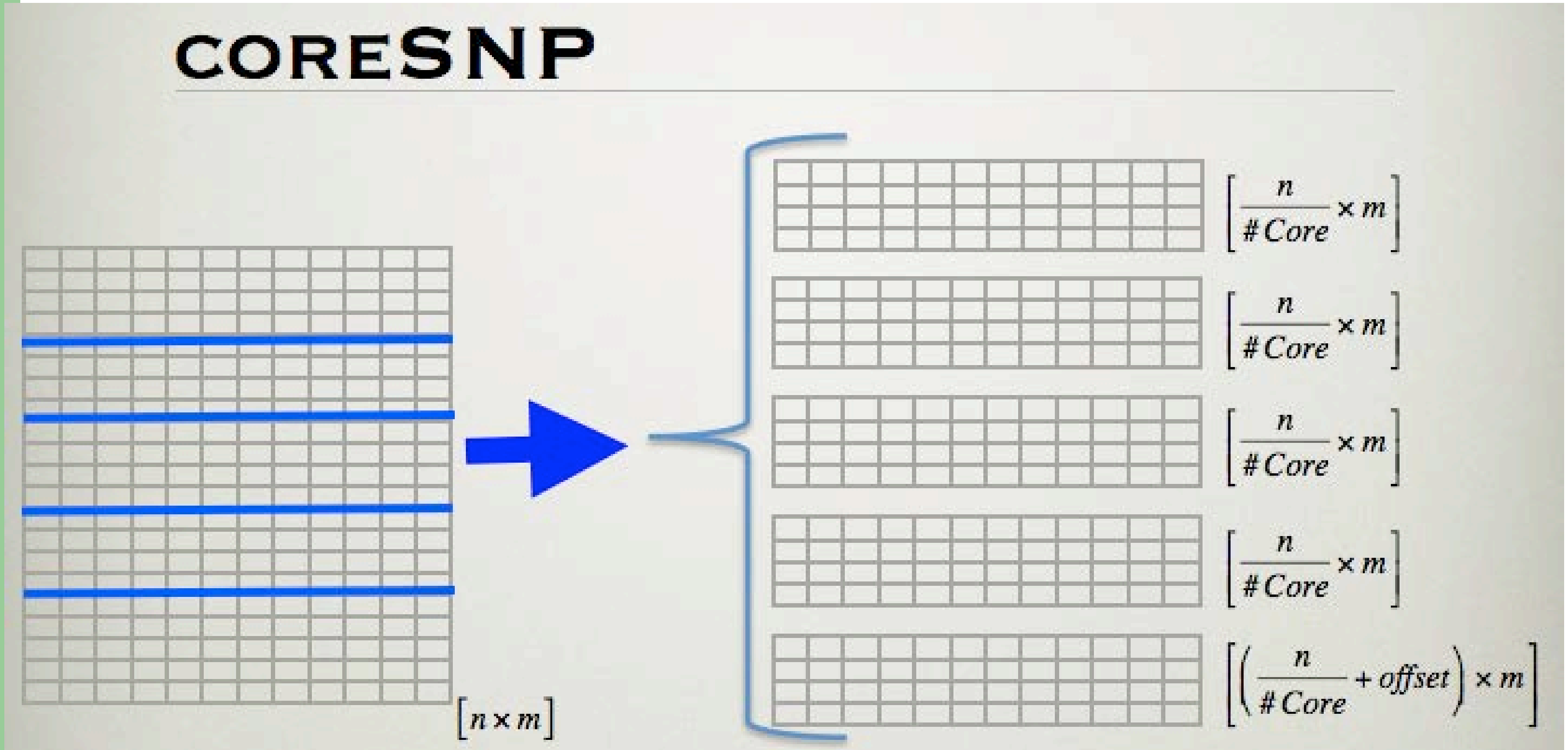
bjh British Journal of Haematology

- A cohort of 19 patients affected by multiple myeloma (MM) treated with aminobisphosphonate zoledronic acid (ZA) were enrolled in a case-control study.
  - 9 patients presented osteonecrosis (ONJ) after the treatment and
  - 10 patients were the control
- The aim of the study was to investigate the association among specific SNPs and the adverse event ONJ induced by ZA.
- Results demonstrated the presence of 8 SNPs that were related to ONJ.
  - Genotypes were determined using DMET Plus GeneChip.
  - Pharmacogenomic profiles were generated by Affymetrix DMET Console software
  - Statistical analysis was performed by two-tailed Fisher’s exact test.
- **DMET-Analyzer identified the 8 SNPs** (with same p-values) that were statistically associated with ONJ occurrence, as those identified in the BJH work using manual analysis, in very less time

coreSNP: the parallel version of DMET-Analyzer

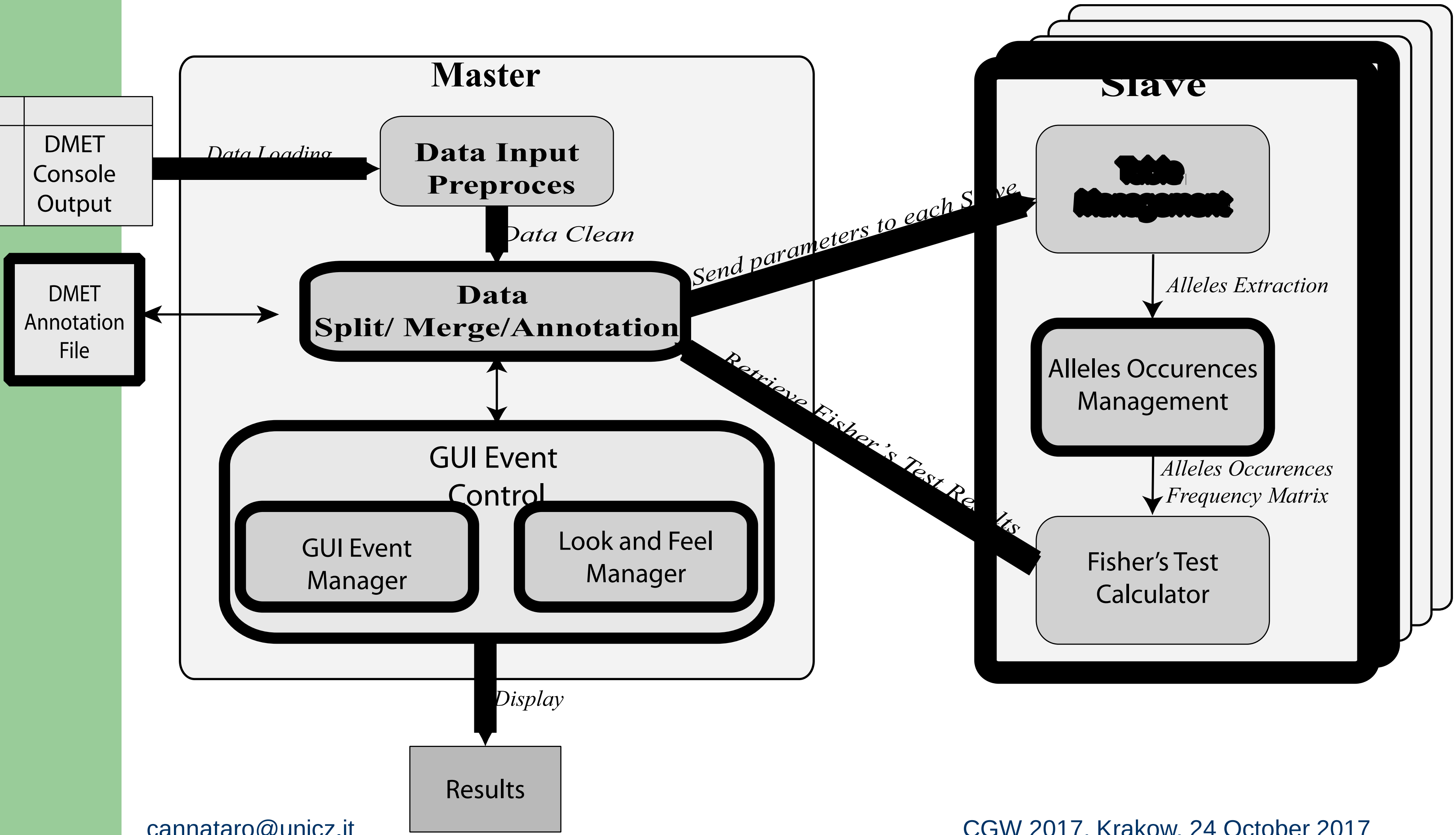
- To face the increasing volume of genotyping data in pharmacogenomics studies, we designed **coreSNP**, a parallel multi-threaded version of DMET-Analyzer
  - Current DMET chips investigate 1936 probes, each one representing a portion of the genome having a role in drug metabolism (225 ADME-related genes are investigated), but **novel chips may investigate millions probes**
  - **Genome Wide Association Studies (GWAS) involve very large populations of patients, thus datasets for pharmacogenomics studies are increasingly huge**
- **coreSNP** takes into account the data parallelism that can be exploited when analyzing DMET data and uses a simple Master/Slave parallel programming approach to decompose computation
- coreSNP uses a **Fisher Significance (Fs)** threshold to discard probes where SNPs distributions among the two classes A and B are very close, i.e. the Fisher tests involving the SNPs detected on that probe are not computed, thus reducing the computational load of the system:
  - A probe *i* is discarded if for each SNP *j*, | *FDT[i,j]* | <= *Fs*, where FDT is a Frequency Difference Table that contains the difference among the frequencies of the SNP j detected on the probe i, respectively in class B and in class A
  - If *Fs=0*, i.e. the most conservative option, only probes with identical alleles distributions in class A and B are discarded.
  - User may also choose to avoid filtering and all probes are tested with Fisher.

Data Parallelism in coreSNP

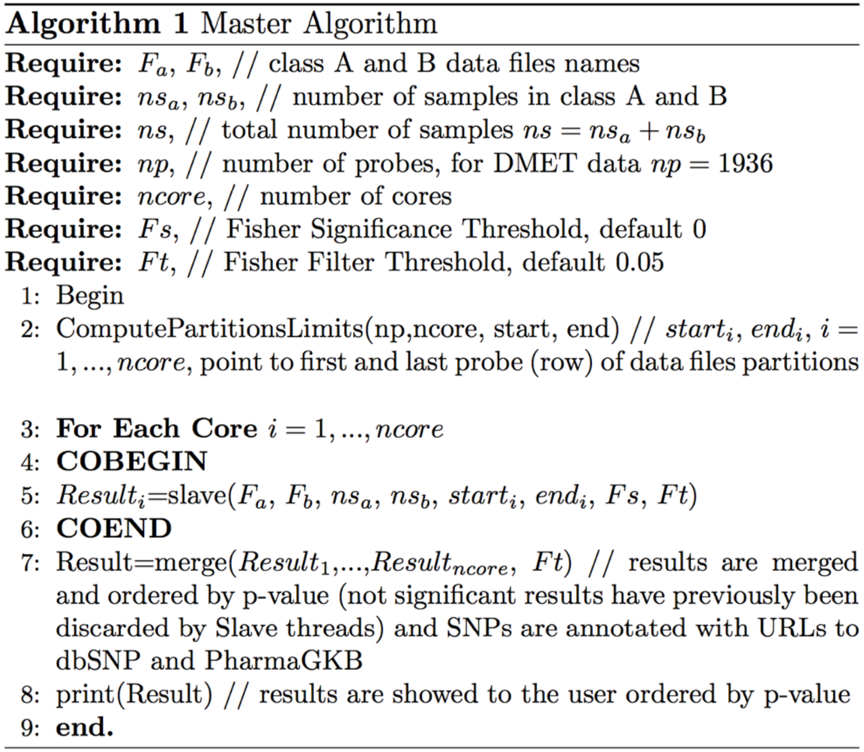

CORESNP
$[n \times m]$
$\left[\frac{n}{\#Core} \times m\right]$
$\left[\frac{n}{\#Core} \times m\right]$
$\left[\frac{n}{\#Core} \times m\right]$
$\left[\frac{n}{\#Core} \times m\right]$
$\left[\left(\frac{n}{\#Core} + offset\right) \times m\right]$

n, number of probes (1936 for current DMET chips)
m, number of samples (patients)
#Core: number of cores (automatically detected by the Master Node)

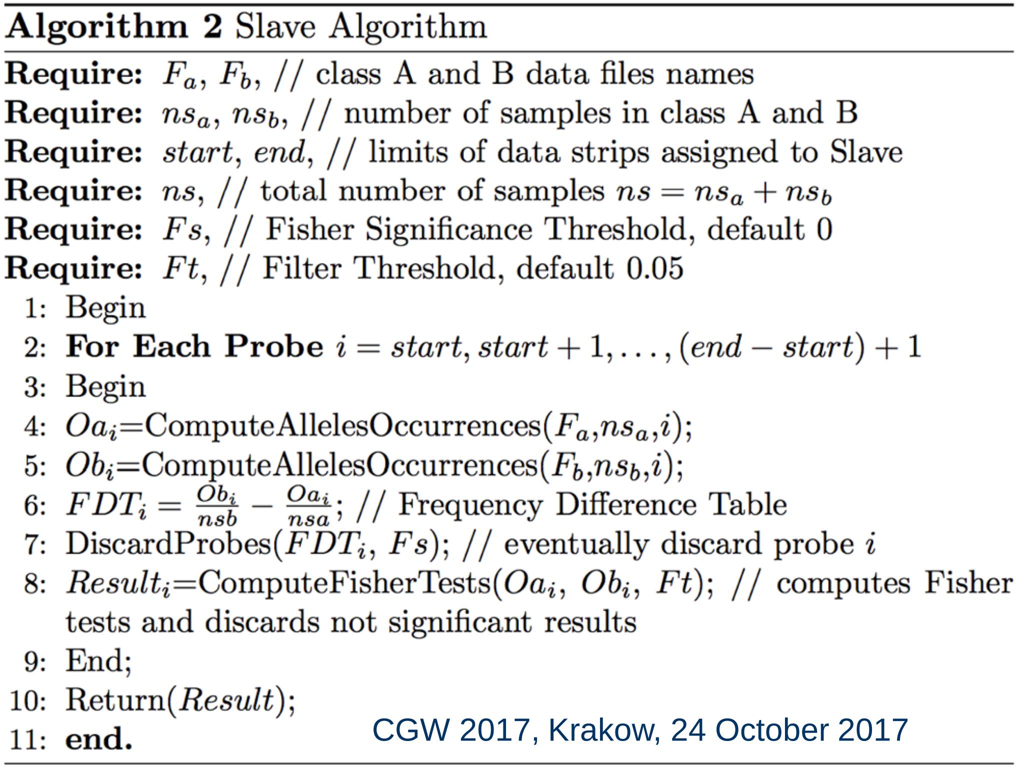
coreSNP Architecture



coreSNP: Master and Slave



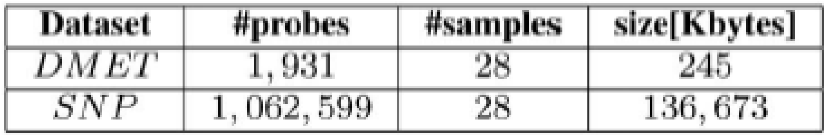
```
Algorithm 1 Master Algorithm
Require: $F_a$, $F_b$, // class A and B data files names
Require: $ns_a$, $ns_b$, // number of samples in class A and B
Require: $ns$, // total number of samples $ns = ns_a + ns_b$
Require: $np$, // number of probes, for DMET data $np = 1936$
Require: $ncore$, // number of cores
Require: $Fs$, // Fisher Significance Threshold, default 0
Require: $Ft$, // Fisher Filter Threshold, default 0.05
Begin
ComputePartitionsLimits(np,ncore, start, end) // $start_i$, $end_i$, $i = 1, ..., ncore$, point to first and last probe (row) of data files partitions
For Each Core $i = 1, ..., ncore$
COBEGIN
$Result_i$=slave($F_a$, $F_b$, $ns_a$, $ns_b$, $start_i$, $end_i$, $Fs$, $Ft$)
COEND
Result=merge($Result_1$,...,$Result_{ncore}$, $Ft$) // results are merged and ordered by p-value (not significant results have previously been discarded by Slave threads) and SNPs are annotated with URLs to dbSNP and PharmaGKB
print(Result) // results are showed to the user ordered by p-value
end.
```



```
Algorithm 2 Slave Algorithm
Require: $F_a$, $F_b$, // class A and B data files names
Require: $ns_a$, $ns_b$, // number of samples in class A and B
Require: $start$, $end$, // limits of data strips assigned to Slave
Require: $ns$, // total number of samples $ns = ns_a + ns_b$
Require: $Fs$, // Fisher Significance Threshold, default 0
Require: $Ft$, // Filter Threshold, default 0.05
Begin
For Each Probe $i = start, start + 1, \ldots, (end - start) + 1$
Begin
$Oa_i$=ComputeAllelesOccurrences($F_a$,$ns_a$,$i$);
$Ob_i$=ComputeAllelesOccurrences($F_b$,$ns_b$,$i$);
$FDT_i = \frac{Ob_i}{nsb} - \frac{Oa_i}{nsa}$; // Frequency Difference Table
DiscardProbes($FDT_i$, $Fs$); // eventually discard probe $i$
$Result_i$=ComputeFisherTests($Oa_i$, $Ob_i$, $Ft$); // computes Fisher tests and discards not significant results
End;
Return($Result$);
end.
```

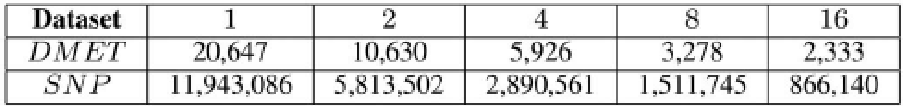
Performance evaluation: SpeedUp

Datasets characteristics; Table Reports the Number of Probes and Samples, and the Datasets Dimension on Disk



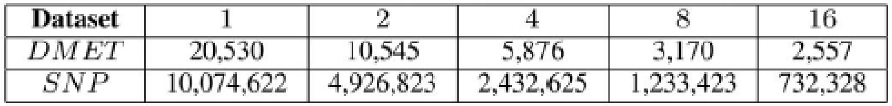
Dataset	#probes	#samples	size[Kbytes]
$DMET$	$1,931$	28	245
SNP	$1,062,599$	28	$136,673$

coreSNP Response Times Without Probe Filtering ($Fs = 0.0$)



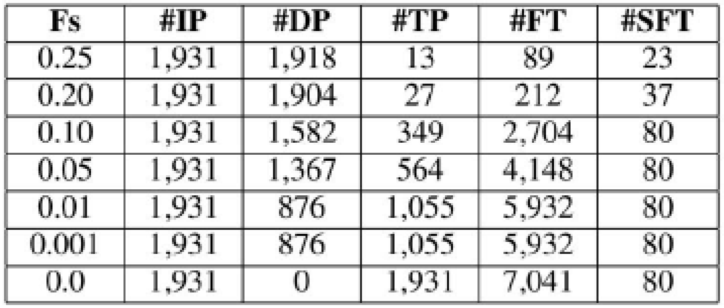
Dataset	1	2	4	8	16
$DMET$	20,647	10,630	5,926	3,278	2,333
SNP	11,943,086	5,813,502	2,890,561	1,511,745	866,140

coreSNP Response Times with a Light Probe Filtering ($Fs = 0.001$)

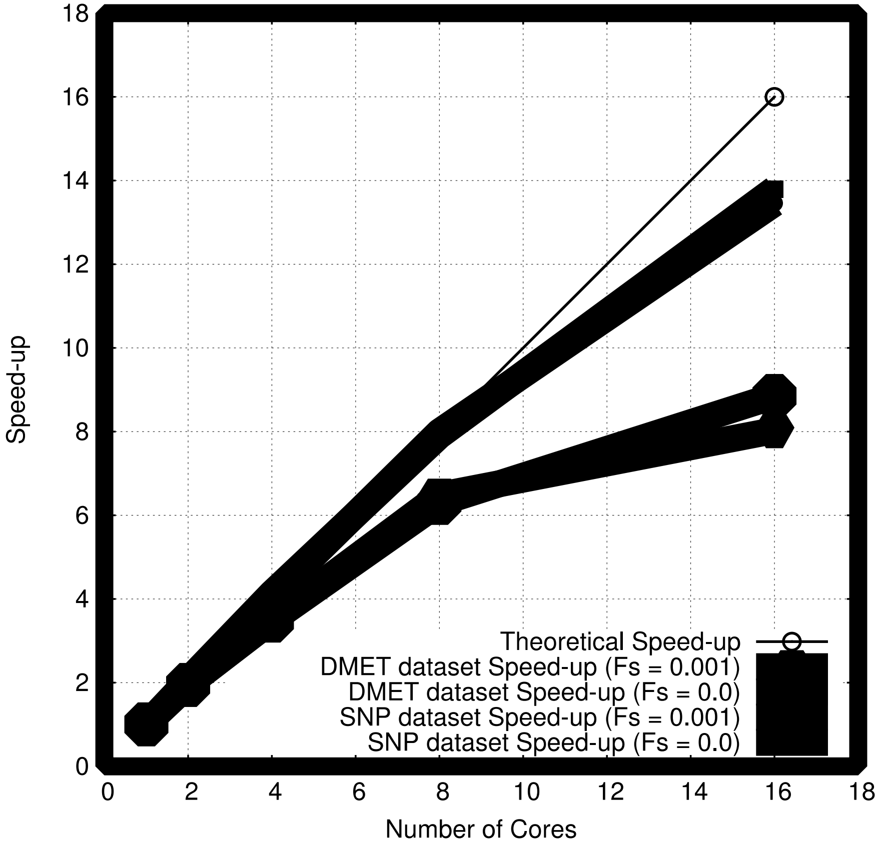


Dataset	1	2	4	8	16
$DMET$	20,530	10,545	5,876	3,170	2,557
SNP	10,074,622	4,926,823	2,432,625	1,233,423	732,328

Input Probes (#IP), Discarded Probes (#DP), Tested Probes (#TP), Executed Fisher Tests (#FT), Significant Fisher Tests (#SFT), when Varying the Fisher Significance Threshold (Fs)



Fs	#IP	#DP	#TP	#FT	#SFT
0.25	1,931	1,918	13	89	23
0.20	1,931	1,904	27	212	37
0.10	1,931	1,582	349	2,704	80
0.05	1,931	1,367	564	4,148	80
0.01	1,931	876	1,055	5,932	80
0.001	1,931	876	1,055	5,932	80
0.0	1,931	0	1,931	7,041	80



Part I: Genomics Data

- Microarray-based Genomics Data
- Automatic preprocessing of gene expression data
  - **micro-CS**: preprocessing and annotation of gene expression data
- Statistical analysis of genotyping data
  - **DMET-Analyzer**: preprocessing and analysis of DMET (Drug Metabolizing Enzymes and Transporters) SNP (Single Nucleotide Polymorphism) data
  - **coreSNP** (multicore implementation)
- Data Mining analysis of genotyping data
  - **DMET-Miner**: Association Rule Mining to extract from DMET data Association Rules able to correlate the contemporary presence of SNPs with patient’s conditions (e.g. TOX vs NOTOX)

Association Rules

- **Association Rules** are used to find frequent patterns, associations, correlations, or causal structures among sets of items in transaction databases.
- **Let X be an item-set, $X \Rightarrow Y$ an association rule and T a set of transactions of a given database**
- **Support** is an indication of how frequently the item-set appears in the database
  - The support value of X with respect T is defined as the proportion of transactions in the database which contains the item-set X, and is indicated as **supp(X)**
- **Confidence** is an indication of how often the rule has been found to be true
  - The confidence value of a rule $X \Rightarrow Y$ with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.
  - Confidence may be interpreted as an estimate of the conditional probability that a transaction having {X} also contains {Y}

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X).$$

- **Goal: Find all rules $X \Rightarrow Y$ with minimum confidence and support**

Mining Association Rules

- Association rules are usually required to satisfy at the same time:
  - a user-specified minimum support and
  - a user-specified minimum confidence.
- Association rule generation is usually split up into two separate steps:
  1. A minimum support threshold is applied to find all frequent item-sets in a database.
  2. A minimum confidence constraint is applied to these frequent item-sets in order to form rules.
- Main algorithms are Apriori and Frequent Pattern-Growth (FP-Growth).

DMET-Miner

- DMET-Miner transforms a case-control DMET dataset in a transaction database and it is able to mine Association Rules from such dataset
- DMET-Miner is based on a modified version of the Frequent Pattern-Growth (FP-Growth) algorithm
- The modified version of FP-Growth needs to scan only twice the database to build an FP-Tree, a structure based on extended prefix-tree, making it possible to store in a compressed way crucial information about the frequent patterns.

DMET-Miner

Journal of Biomedical Informatics

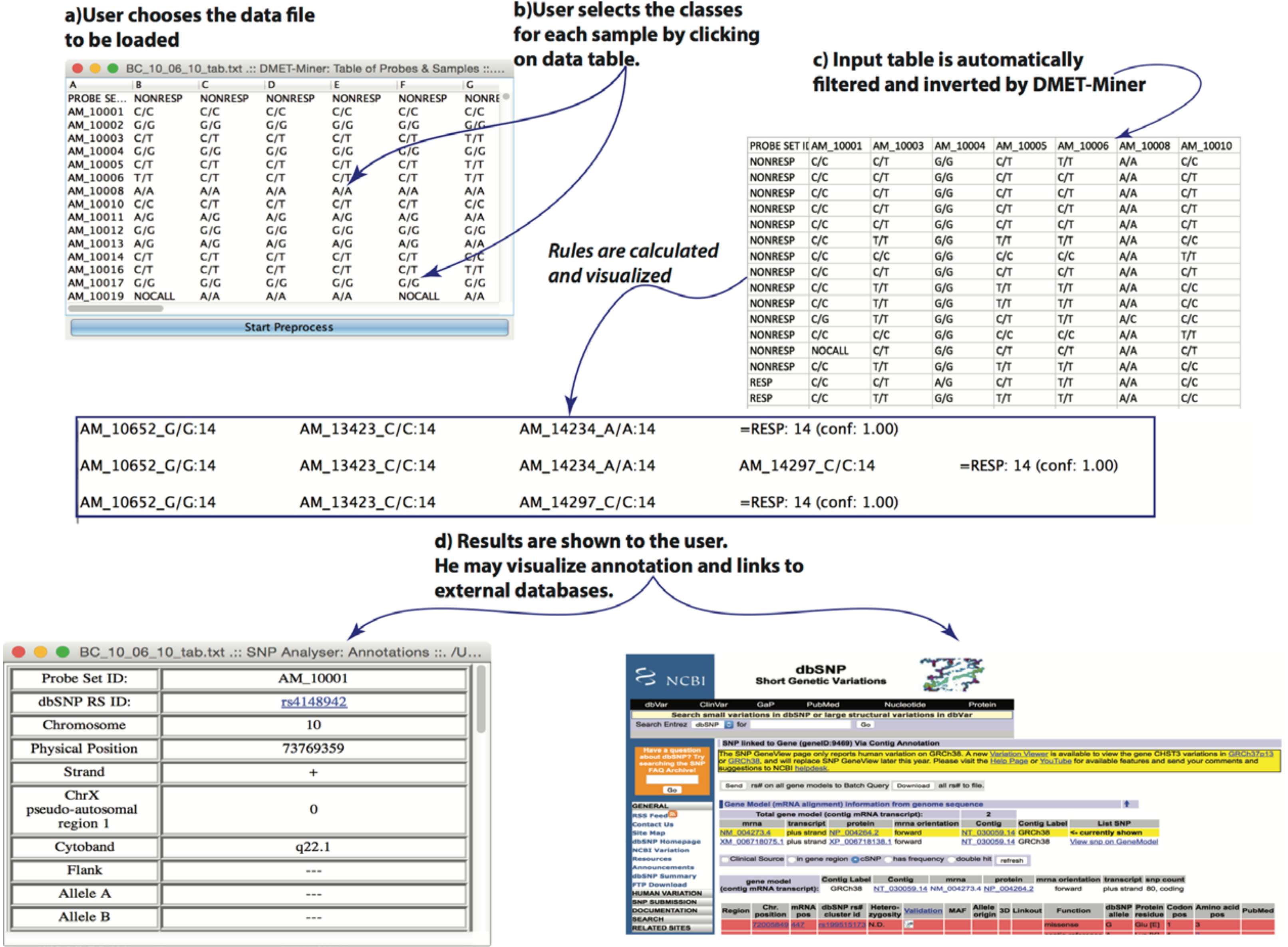
Volume 56, August 2015, Pages 273–283

DMET-Miner: Efficient discovery of association rules from pharmacogenomic data

Giuseppe Agapito[a], Pietro H. Guzzi[a], Mario Cannataro[a, b]

⊞ Show more

doi:10.1016/j.jbi.2015.06.005



a)User chooses the data file to be loaded

BC_10_06_10_tab.txt .:: DMET-Miner: Table of Probes & Samples ::....

A	B	C	D	E	F	G
PROBE SE...	NONRESP	NONRESP	NONRESP	NONRESP	NONRESP	NONRE
AM_10001	C/C	C/C	C/C	C/C	C/C	C/C
AM_10002	G/G	G/G	G/G	G/G	G/G	G/G
AM_10003	C/T	C/T	C/T	C/T	C/T	T/T
AM_10004	G/G	G/G	G/G	G/G	G/G	G/G
AM_10005	C/T	C/T	C/T	C/T	C/T	T/T
AM_10006	T/T	C/T	C/T	C/T	C/T	T/T
AM_10008	A/A	A/A	A/A	A/A	A/A	A/A
AM_10010	C/C	C/T	C/T	C/T	C/T	C/C
AM_10011	A/G	A/G	A/G	A/G	A/G	A/A
AM_10012	G/G	G/G	G/G	G/G	G/G	G/G
AM_10013	A/G	A/G	A/G	A/G	A/G	A/A
AM_10014	C/T	C/T	C/T	C/T	C/T	C/C
AM_10016	C/T	C/T	C/T	C/T	C/T	T/T
AM_10017	G/G	G/G	G/G	G/G	G/G	G/G
AM_10019	NOCALL	A/A	A/A	A/A	NOCALL	A/A

b)User selects the classes for each sample by clicking on data table.

c) Input table is automatically filtered and inverted by DMET-Miner

PROBE SET II	AM_10001	AM_10003	AM_10004	AM_10005	AM_10006	AM_10008	AM_10010
NONRESP	C/C	C/T	G/G	C/T	T/T	A/A	C/C
NONRESP	C/C	C/T	G/G	C/T	C/T	A/A	C/T
NONRESP	C/C	C/T	G/G	C/T	C/T	A/A	C/T
NONRESP	C/C	C/T	G/G	C/T	C/T	A/A	C/T
NONRESP	C/C	C/T	G/G	C/T	C/T	A/A	C/T
NONRESP	C/C	T/T	G/G	T/T	T/T	A/A	C/C
NONRESP	C/C	C/C	G/G	C/C	C/C	A/A	T/T
NONRESP	C/C	C/T	G/G	C/T	C/T	A/A	C/T
NONRESP	C/C	T/T	G/G	T/T	T/T	A/A	C/C
NONRESP	C/C	T/T	G/G	T/T	T/T	A/A	C/C
NONRESP	C/G	T/T	G/G	C/T	T/T	A/C	C/C
NONRESP	C/C	C/C	G/G	C/C	C/C	A/A	T/T
NONRESP	NOCALL	C/T	G/G	C/T	C/T	A/A	C/T
NONRESP	C/C	T/T	G/G	T/T	T/T	A/A	C/C
RESP	C/C	C/T	A/G	C/T	T/T	A/A	C/C
RESP	C/C	T/T	G/G	T/T	T/T	A/A	C/C

Rules are calculated and visualized

```
AM_10652_G/G:14    AM_13423_C/C:14    AM_14234_A/A:14    =RESP: 14 (conf: 1.00)

AM_10652_G/G:14    AM_13423_C/C:14    AM_14234_A/A:14    AM_14297_C/C:14    =RESP: 14 (conf: 1.00)

AM_10652_G/G:14    AM_13423_C/C:14    AM_14297_C/C:14    =RESP: 14 (conf: 1.00)
```

d) Results are shown to the user. He may visualize annotation and links to external databases.

BC_10_06_10_tab.txt .:: SNP Analyser: Annotations ::. /U...

Probe Set ID:	AM_10001
dbSNP RS ID:	rs4148942
Chromosome	10
Physical Position	73769359
Strand	+
ChrX pseudo-autosomal region 1	0
Cytoband	q22.1
Flank	---
Allele A	---
Allele B	---

SNP linked to Gene (geneID:9469) Via Contig Annotation

The SNP GeneView page only reports human variation on GRCh38. A new Variation Viewer is available to view the gene CHST3 variations in GRCh37p13 or GRCh38, and will replace SNP GeneView later this year. Please visit the Help Page or YouTube for available features and send your comments and suggestions to NCBI helpdesk.

Send rs# on all gene models to Batch Query Download all rs# to file.

Gene Model (mRNA alignment) information from genome sequence

Total gene model (contig mRNA transcript): 2

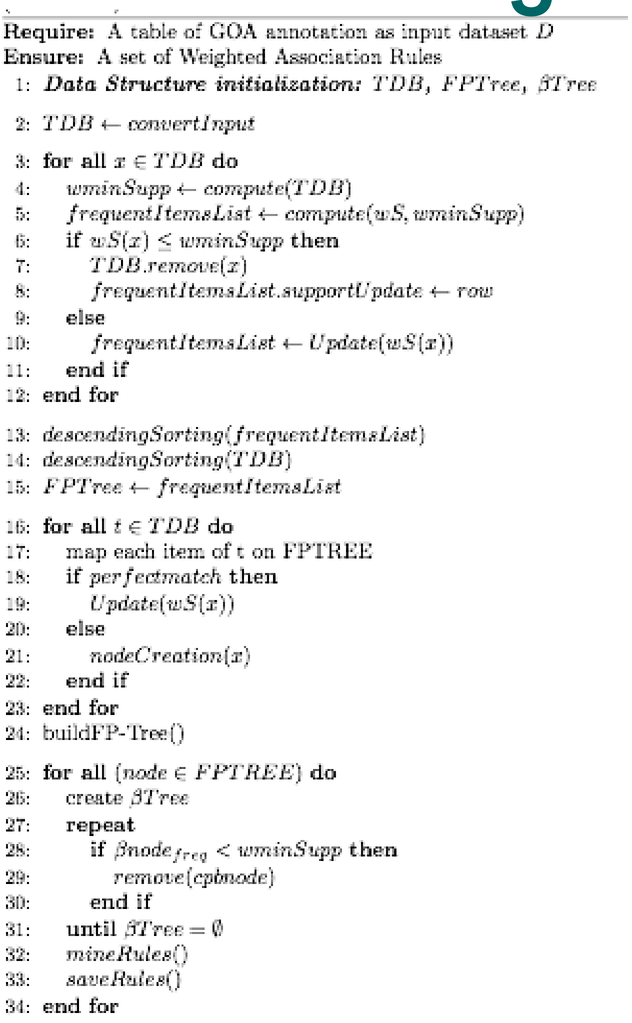
mrna	transcript	protein	mrna orientation	Contig	Contig Label	List SNP
NM_004273.4	plus strand	NP_004264.2	forward	NT_030059.14	GRCh38	<- currently shown
XM_006718075.1	plus strand	XP_006718138.1	forward	NT_030059.14	GRCh38	View snp on GeneModel

□ Clinical Source ○ in gene region ● cSNP ○ has frequency ○ double hit refresh

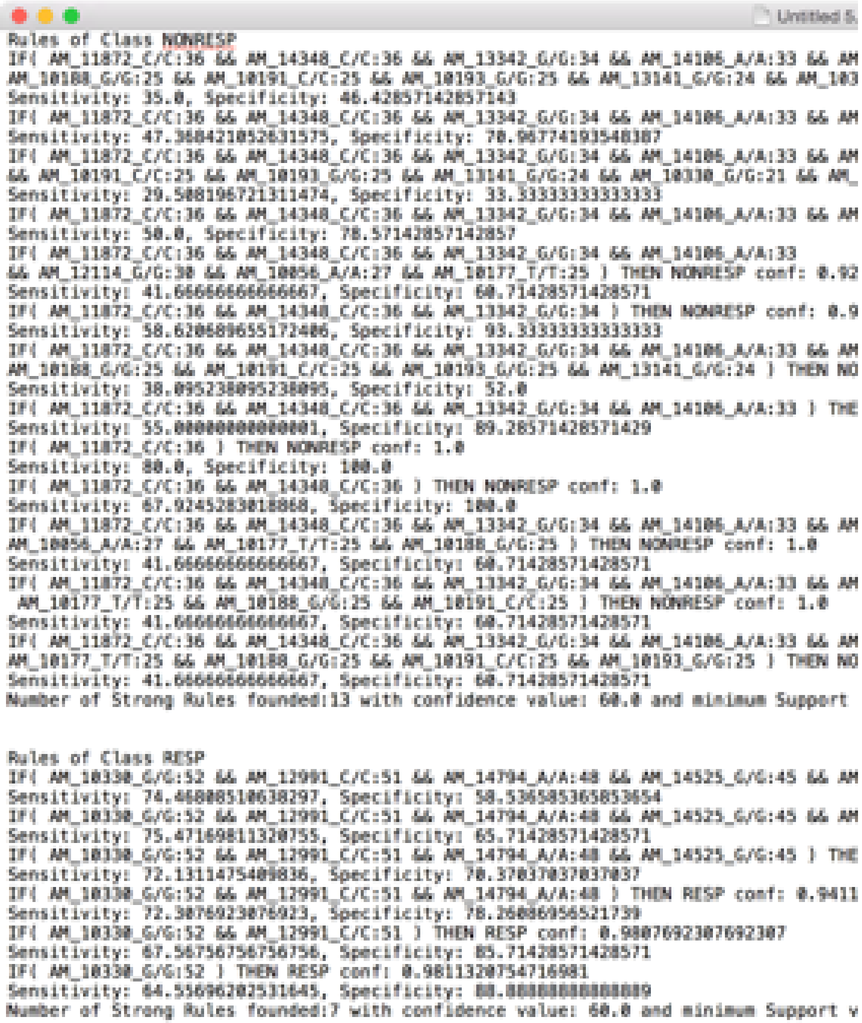
gene model (contig mRNA transcript):	Contig Label	Contig	mrna	protein	mrna orientation	transcript	snp count
	GRCh38	NT_030059.14	NM_004273.4	NP_004264.2	forward	plus strand	80, coding

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	MAF	Allele origin	3D	Linkout	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed
	72005849	447	rs199515173	N.D.						missense	G	Glu [E]	1	3	

DMET-Miner Algorithm and Mined Rules



```
Require: A table of GOA annotation as input dataset $D$
Ensure: A set of Weighted Association Rules
Data Structure initialization: $TDB$, $FPTree$, $\beta Tree$
$TDB \leftarrow convertInput$
for all $x \in TDB$ do
    $wminSupp \leftarrow compute(TDB)$
    $frequentItemsList \leftarrow compute(wS, wminSupp)$
    if $wS(x) \leq wminSupp$ then
        $TDB.remove(x)$
        $frequentItemsList.supportUpdate \leftarrow row$
    else
        $frequentItemsList \leftarrow Update(wS(x))$
    end if
end for
$descendingSorting(frequentItemsList)$
$descendingSorting(TDB)$
$FPTree \leftarrow frequentItemsList$
for all $t \in TDB$ do
    map each item of t on FPTREE
    if $perfectmatch$ then
        $Update(wS(x))$
    else
        $nodeCreation(x)$
    end if
end for
buildFP-Tree()
for all $(node \in FPTREE)$ do
    create $\beta Tree$
    repeat
        if $\beta node_{freq} < wminSupp$ then
            $remove(cpbnode)$
        end if
    until $\beta Tree = \emptyset$
    $mineRules()$
    $saveRules()$
end for
```



```
Rules of Class NONRESP
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
AM_10188_G/G:25 && AM_10191_C/C:25 && AM_10193_G/G:25 && AM_13141_G/G:24 && AM_103
Sensitivity: 35.0, Specificity: 46.42857142857143
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
Sensitivity: 47.368421052631575, Specificity: 70.96774193548387
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
&& AM_10191_C/C:25 && AM_10193_G/G:25 && AM_13141_G/G:24 && AM_10330_G/G:21 && AM_
Sensitivity: 29.508196721311474, Specificity: 33.333333333333333
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
Sensitivity: 50.0, Specificity: 78.57142857142857
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33
&& AM_12114_G/G:30 && AM_10056_A/A:27 && AM_10177_T/T:25 ) THEN NONRESP conf: 0.92
Sensitivity: 41.666666666666667, Specificity: 60.71428571428571
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 ) THEN NONRESP conf: 0.9
Sensitivity: 58.620689655172406, Specificity: 93.333333333333333
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
AM_10188_G/G:25 && AM_10191_C/C:25 && AM_10193_G/G:25 && AM_13141_G/G:24 ) THEN NO
Sensitivity: 38.095238095238095, Specificity: 52.0
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 ) THE
Sensitivity: 55.00000000000001, Specificity: 89.28571428571429
IF( AM_11872_C/C:36 ) THEN NONRESP conf: 1.0
Sensitivity: 80.0, Specificity: 100.0
IF( AM_11872_C/C:36 && AM_14348_C/C:36 ) THEN NONRESP conf: 1.0
Sensitivity: 67.9245283018868, Specificity: 100.0
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
AM_10056_A/A:27 && AM_10177_T/T:25 && AM_10188_G/G:25 ) THEN NONRESP conf: 1.0
Sensitivity: 41.666666666666667, Specificity: 60.71428571428571
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
 AM_10177_T/T:25 && AM_10188_G/G:25 && AM_10191_C/C:25 ) THEN NONRESP conf: 1.0
Sensitivity: 41.666666666666667, Specificity: 60.71428571428571
IF( AM_11872_C/C:36 && AM_14348_C/C:36 && AM_13342_G/G:34 && AM_14106_A/A:33 && AM
AM_10177_T/T:25 && AM_10188_G/G:25 && AM_10191_C/C:25 && AM_10193_G/G:25 ) THEN NO
Sensitivity: 41.666666666666667, Specificity: 60.71428571428571
Number of Strong Rules founded:13 with confidence value: 60.0 and minimum Support


Rules of Class RESP
IF( AM_10330_G/G:52 && AM_12991_C/C:51 && AM_14794_A/A:48 && AM_14525_G/G:45 && AM
Sensitivity: 74.46808510638297, Specificity: 58.536585365853654
IF( AM_10330_G/G:52 && AM_12991_C/C:51 && AM_14794_A/A:48 && AM_14525_G/G:45 && AM
Sensitivity: 75.47169811320755, Specificity: 65.71428571428571
IF( AM_10330_G/G:52 && AM_12991_C/C:51 && AM_14794_A/A:48 && AM_14525_G/G:45 ) THE
Sensitivity: 72.1311475409836, Specificity: 70.37037037037037
IF( AM_10330_G/G:52 && AM_12991_C/C:51 && AM_14794_A/A:48 ) THEN RESP conf: 0.9411
Sensitivity: 72.3076923076923, Specificity: 78.26086956521739
IF( AM_10330_G/G:52 && AM_12991_C/C:51 ) THEN RESP conf: 0.9807692307692307
Sensitivity: 67.56756756756756, Specificity: 85.71428571428571
IF( AM_10330_G/G:52 ) THEN RESP conf: 0.9811320754716981
Sensitivity: 64.55696202531645, Specificity: 88.88888888888889
Number of Strong Rules founded:7 with confidence value: 60.0 and minimum Support v
```

- DMET-Miner produces rules able to discriminate distinctive features for each class.
- DMET-Miner makes easy to profile in what class new subjects belong (e.g. RESP, NORESP) on the basis of their SNPs

Comparison among DMET-Miner, Weka and Rapid Miner

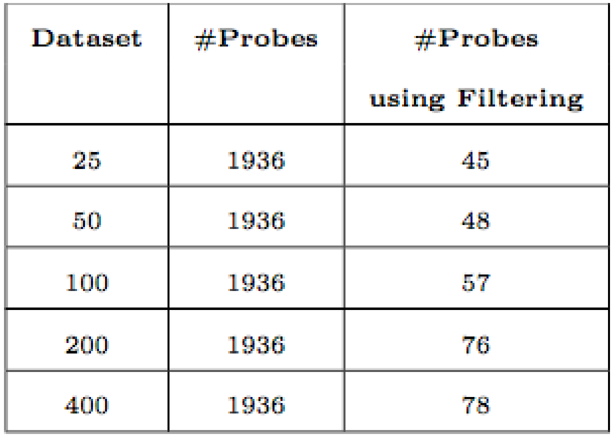
- We have compared the FP-Growth algorithm implemented in DMET-Miner with the FP-Growth algorithm implemented in Weka (version 3.6.10) and the FP-Growth algorithm available in RapidMiner (version 5.3.013).
- Weka and RapidMiner were not able to directly load the DMET dataset produced by the DMET platform
- In order to compare DMET-Miner FP-Growth with Weka FP-Growth and RapidMiner FP-Growth on the same conditions, we have given as input to Weka and RapidMiner the filtered dataset produced by our software.
- In this way we ensure that the inputs given to DMET-Miner, RapidMiner and Weka have the same dimensions.

Datasets

- The comparison of the FP-Growth algorithm implemented in the three software tools was tested using five synthetic DMET datasets. We built the synthetic datasets containing the same number of probes as a real DMET dataset (1, 936 probes) and doubling the number of samples (in the 13 experiments we analyzed five datasets with respectively 25, 50, 100, 200 and 400 samples for each dataset) grouped into two classes.
- We populated these datasets with randomly significantly different distributions of SNPs.
- The datasets contain data related to samples from subjects belonging to two classes: subjects which respond to drugs (class RESP) and subjects which do not respond to drugs (class NONRESP), simulating a classical case-control study.
- The dimensions of the datasets analyzed range from 300KB for the dataset with 25 samples to about 3.1MB for the one with 400 samples

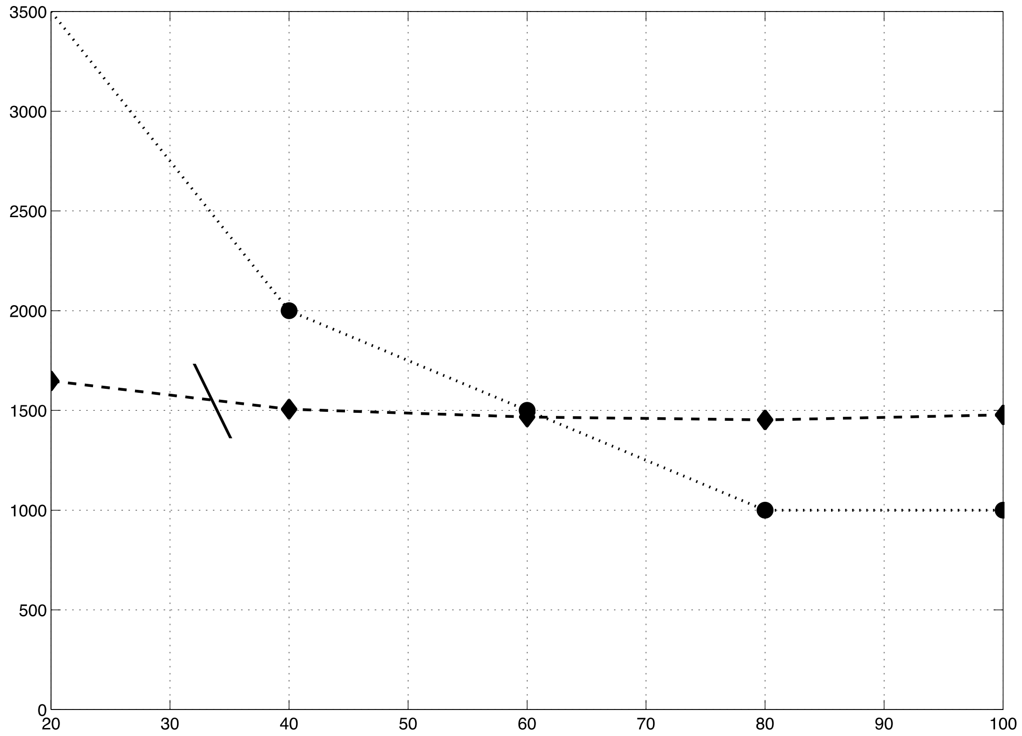
Fisher Test Filtering

- Number of meaningful probes (rows) after using the Fisher Test Filtering



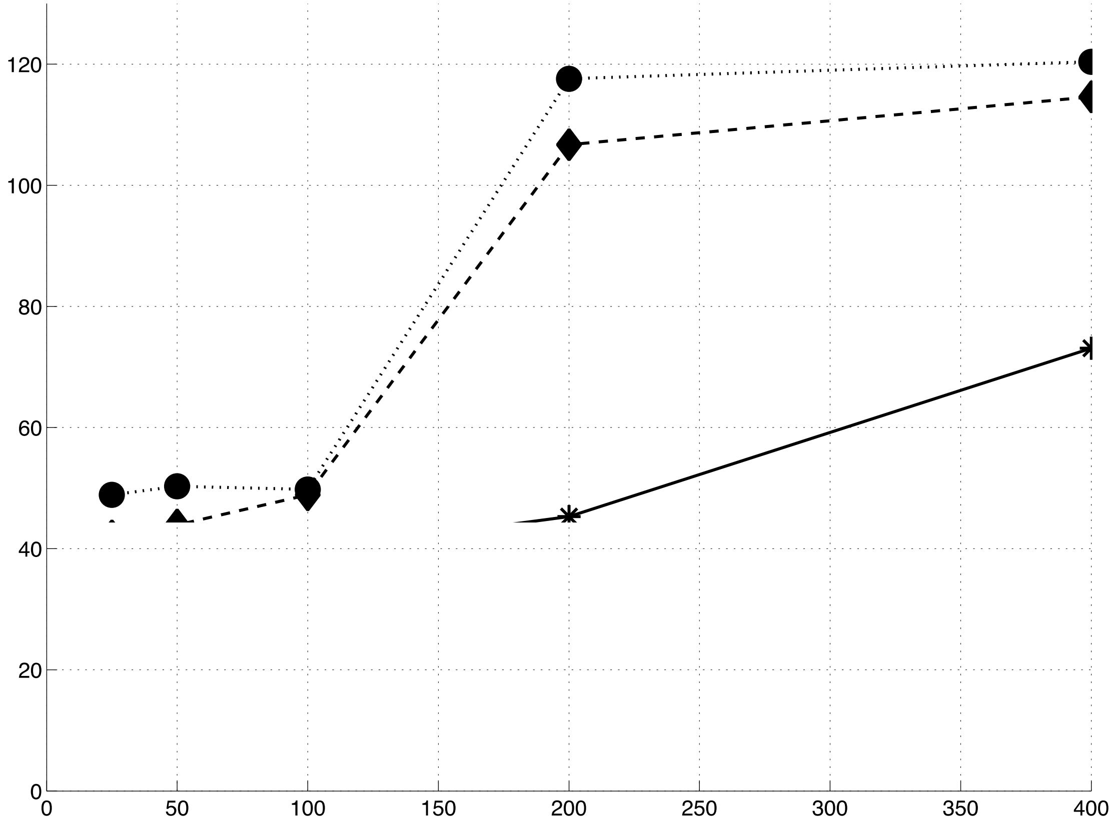
Dataset	#Probes	#Probes using Filtering
25	1936	45
50	1936	48
100	1936	57
200	1936	76
400	1936	78

Execution Times



Execution time of the algorithms varying the minimum support using the 100 samples dataset. The execution time is obtained in function of the value of minimum support used and the number of mined rules. The dotted line (top part of figure) represents the running time of RapidMiner, the dashed line (middle part of figure), represents running time of Weka, the continuous line (bottom part of figure) represents the running time of DMET-Miner.

Memory Consumption



Memory consumption of RapidMiner, Weka and DMET-Miner when doubling the size of the input dataset. The dotted line (top part of figure) represents the memory consumption of RapidMiner, the dashed line (middle part of figure), represents the memory consumption time of Weka, the continuous line (bottom part of figure) represents the memory consumption of DMET-Miner.

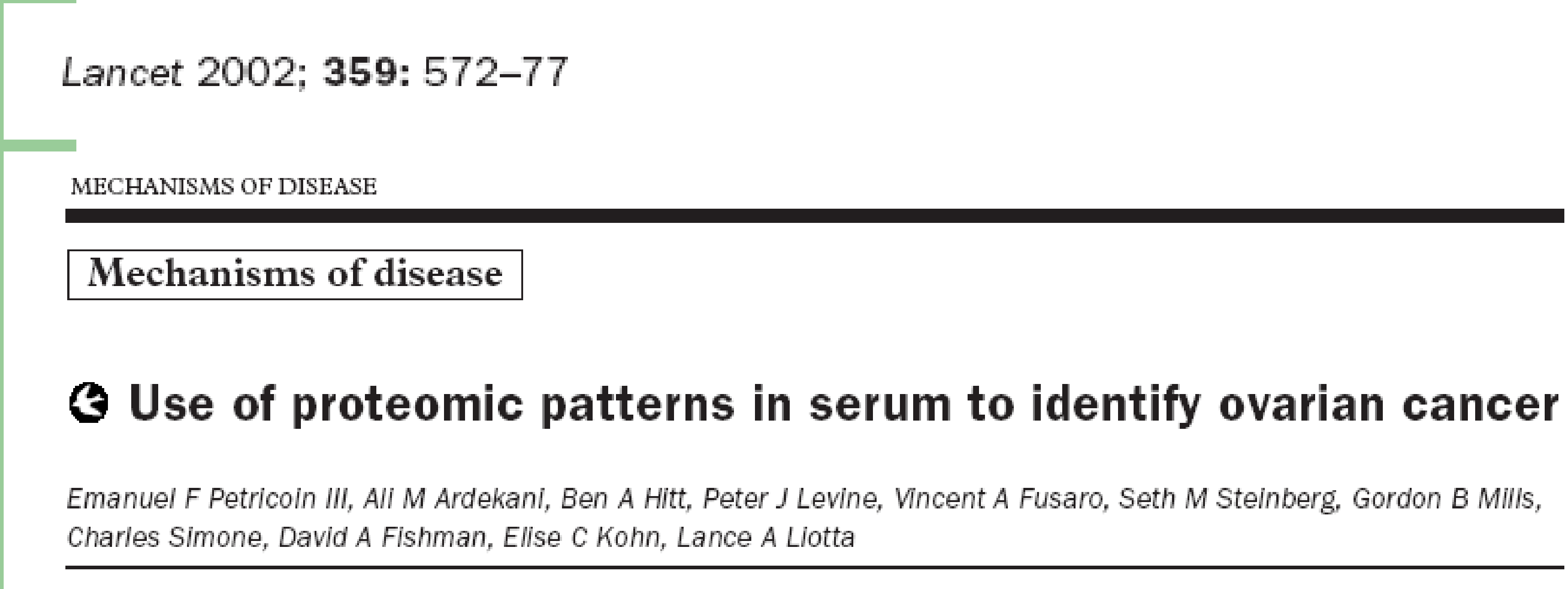
OUTLINE

- Experiences at University of Catanzaro in high performance management, preprocessing and analysis of omics data
  - PART I: Genomics data
  - **PART II: Proteomics data**
  - PART III: Interactomics data
- Conclusions

Proteomics data

- Introduction to Mass Spectrometry
- MS-Analyzer: service oriented platform for preprocessing and mining of MS data
  - MS-Analyzer = Ontology + Spectra Services + Workflow
- EIPEPTIDI: enhanced protein identification from ICAT MS/MS data

Early diagnosis of cancer



Lancet 2002; **359:** 572–77

MECHANISMS OF DISEASE

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

Recently MS has been used to analyse low weight proteins present in serum/plasma for early diagnosis of tumours

UMG

MS-based Proteomics

Sample introduction

Interface to Vacuum

Ion Source

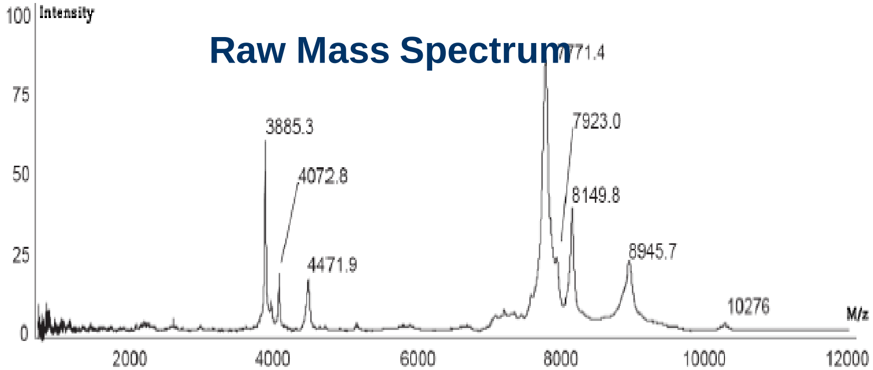
Mass Analyser

Detector

High Vacuum

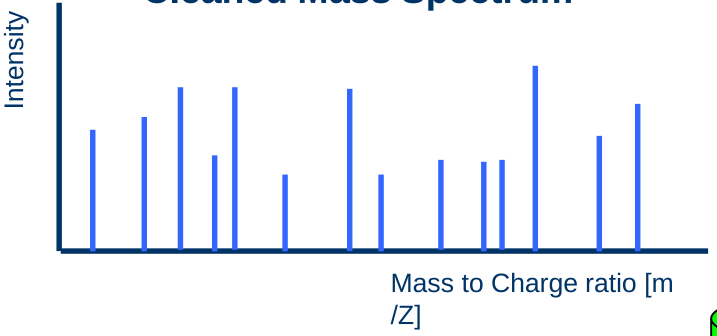
Electron Impact Ionisation

Mass Spectrometer



Raw Mass Spectrum

Cleaned Mass Spectrum



Data Preparation

Workflow Execution (i.e. Supervised Classification)

Knowledge Models

Data Pre-Processing

Subtract Base Line
Normalization
Binning
Peak Identification
Peak Alignment

Preprocessed Cleaned Spectra Database

MS at UMG

MALDI-TOF

Applied Biosystems Voyager DE-STR

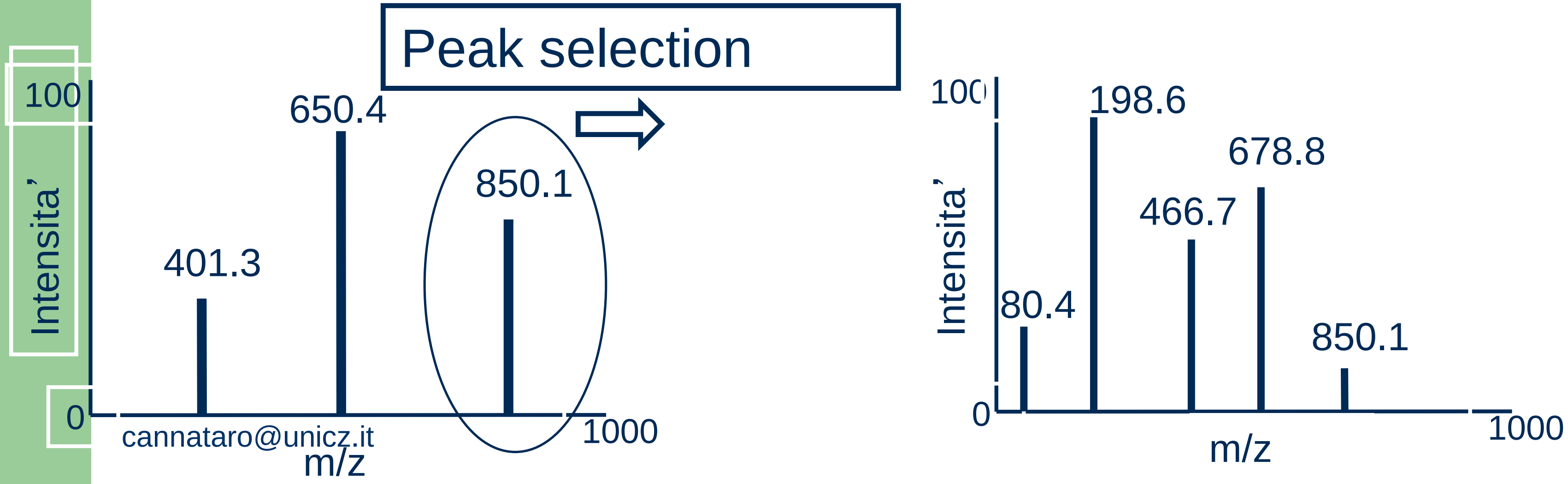
ESI-QqTOF

Applied Biosystems QSTAR XL LC-MS/MS

Tandem mass spectrometry (MS/MS)

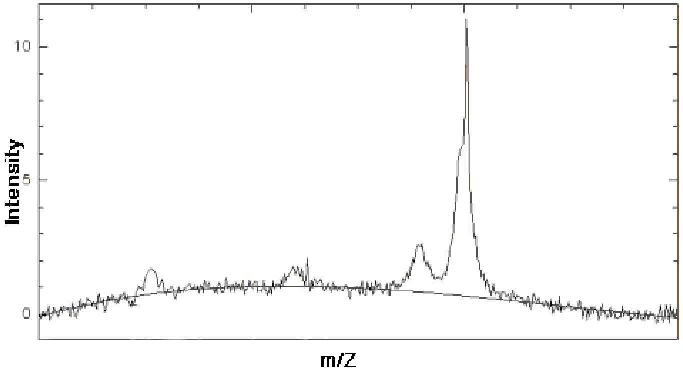
MS1: MS spectrum
from 0 to 1000 m/z (Da)

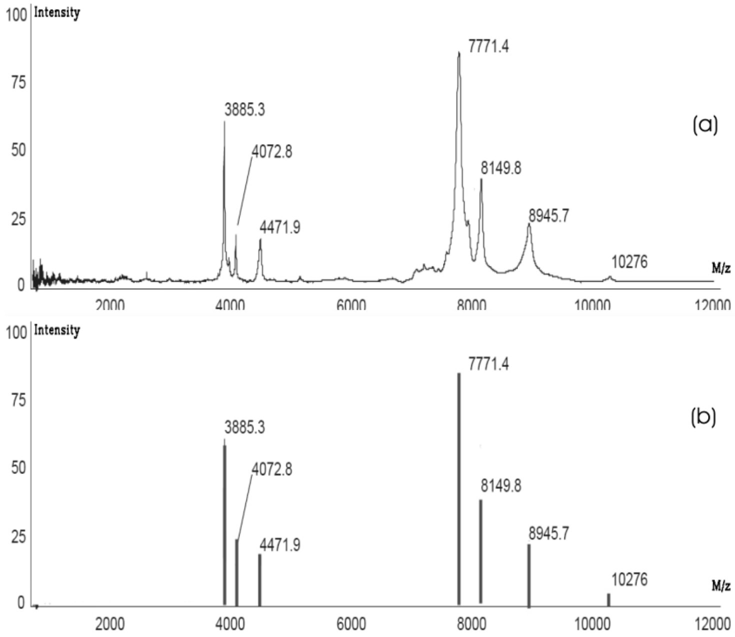
MS2: MS/MS
on m/z 850.1



Spectra Preprocessing

- ***Noise Reduction***
  - ***Base line subtraction*** flattens the base profile of a spectrum
  - ***Smoothing*** reduces the noise level in the whole spectrum.
- ***Binning***
  - groups measured data into bins
    - aggregate intensity (e.g. the sum of the intensities in the bin)
    - representative m/Z value (e.g. the median or the one with maximum intensity)
- ***Normalization***
  - aims to make intensity comparable across different spectra.
- ***Peaks alignment***
  - The same peak (e.g. the same peptide) may have different m/Z values across samples
  - finds a common set of peak locations (i.e. m/Z values) in a set of spectra, so that all spectra have common m/Z values for the same biological entities





Spectra analysis

- Biomarker discovery
  - Analysis of collection of spectra (usually by using data mining or machine learning) to find discriminating peaks among different conditions (e.g. healthy, diseased)
- Protein/Peptide identification
  - Identification of protein/peptide associated to a peak, e.g. through MS/MS and database search (Mascot, Sequest, X!Tandem)
- Qualitative vs quantitative proteomics
  - MALDI-TOF or SELDI-TOF give a snapshot of a sample, without providing quantitative information nor complete identification
  - Labelling techniques, such as ICAT or SILAC, coupled to tandem MS (E.g. ICAT-based LC-MS/MS) allow quantification and identification of proteins

MS data

- The basic MS data is a long sequence of (intensity, m/z) value pairs
- Indeed, different MS data formats do exist depending on:
  - Ionization (ESI, MALDI, TOF)
  - Separation (e.g. GAS or Liquid Chromatography)
    - More spectra at different (retention) times
  - MS approach (e.g. MS vs MS/MS)
    - One spectrum/sample vs more spectra/sample
  - Labeling
    - intensity in a sample is a measure relative to a control one
- Moreover, each MS company usually uses a proprietary format
- Heterogeneity in spectra data is an issue

S. Orchard, H. Hermjakob, and R. Apweiler. The proteomics standards initiative. Proteomics, 3(7):1374.1376, 2003.

mzData

- mzData was the first XML-based data model defined by HUPO-PSI (Human Proteome Organization-Proteomics Standard Initiative) to standardize mass spectrometry-based experimental data
- mzData schema comprises a
  - **description** element describing metadata about the experiment (e.g. administrative information, data about instrument and software used to generate the spectra),
  - and a **spectrumList** element that encloses a set of **spectrum** elements.
- Each spectrum element stores, respectively, all the **m/z** and **intensities** values of the spectrum as Base64 encoded strings.
  - Base64 encoding allows to represent arbitrary sequences of octets through a 65-character alphabet (each 6 bits are represented as a printable character).
  - Base64 allows to easily transmit MS data over Internet protocols and can be stored into XML documents.

Issues in Mining Spectra Data

- **Loading** of the raw spectra produced by mass spectrometer
- **Converting** (eventually) in a standard format (mzXML, mzData),
- **Preprocessing** of the raw spectra data,
- **Preparation** of the data mining input file (e.g. Weka ARFF file),
- **Data Mining analysis** (e.g. classification) of mass spectra,
  - CRISP-DM methodology,
  - Clementine, Weka, and SPSS platforms
- **Knowledge Models visualization** (e.g. decision tree)
  - represented in a standard language as PMML

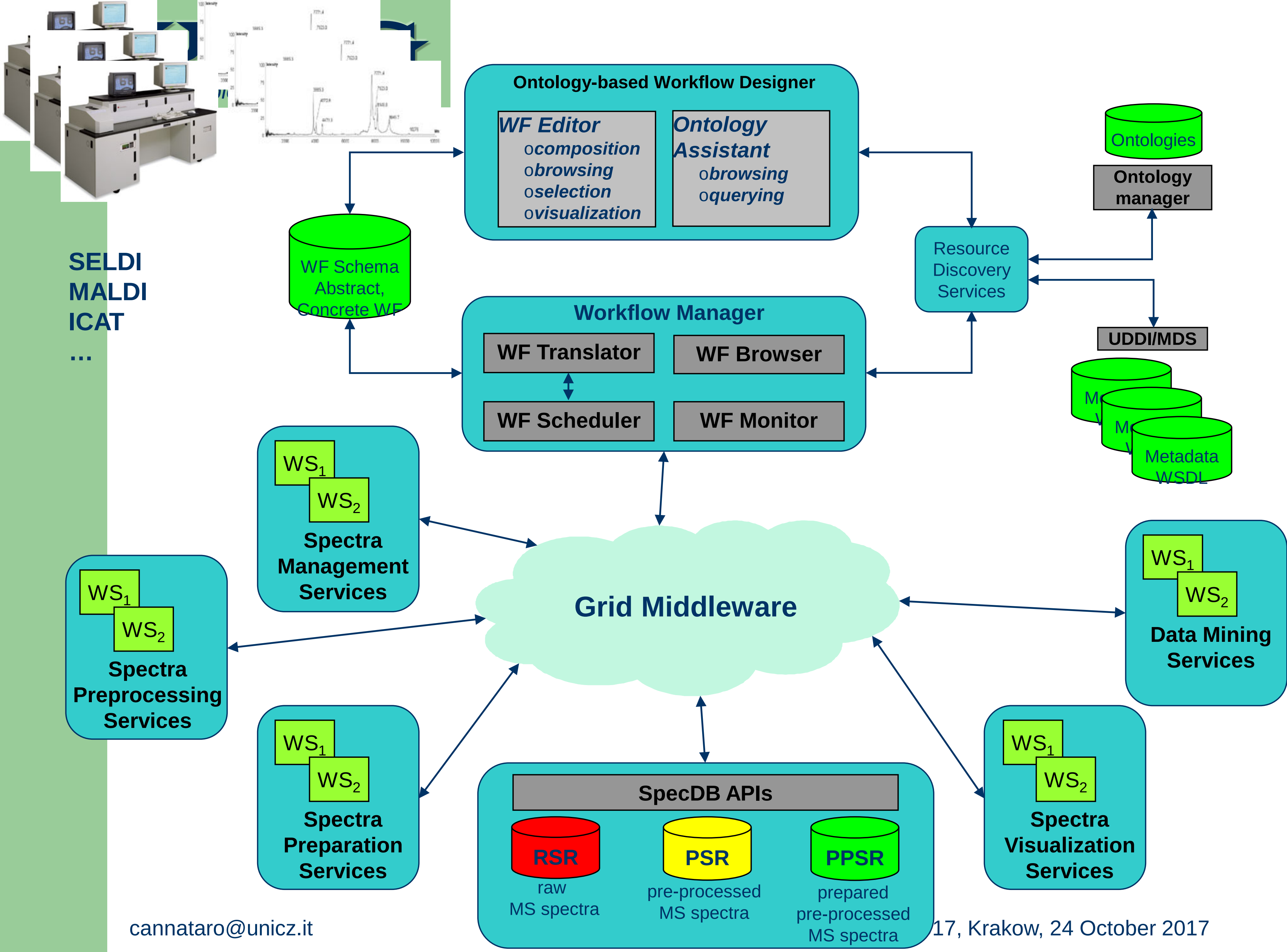
- **Key issues**
  - Service Oriented Distributed Computation
  - Efficient spectra management
  - Provenance data

Cannataro M, Barla A, Flor R, Jurman G, Merler S, Paoli S, Tradigo G, Veltri P, Furlanello C. A grid environment for high-throughput proteomics. **IEEE Trans Nanobioscience**, 6(2):117-23, 2007.

M. Cannataro, P. Veltri: MS-Analyzer: preprocessing and data mining services for proteomics applications on the Grid. **Concurrency and Computation: Practice and Experience** 19(15): 2047-2066, 2007.

MS-Analyzer

- MS-Analyzer is a software platform for the preprocessing, management and data mining analysis of MS data. It uses
  - domain ontologies to model
    - software tools (e.g. preprocessing and mining) and their relationships
    - data sources (e.g. MALDI-TOF, LC-MS/MS spectra datasets)
    - constraints (e.g. binning cannot be applied twice)
  - and workflow techniques to design “in silico” experiments.
- MS-Analyzer:
  - acquires, preprocesses and manages MS data
  - offers filtering, preprocessing, and DM services
  - sharing experiments data, workflows and knowledge



SELDI
MALDI
ICAT
…

Ontology-based Workflow Designer

WF Editor
- o*composition*
- o*browsing*
- o*selection*
- o*visualization*

Ontology Assistant
- o*browsing*
- o*querying*

WF Schema Abstract, Concrete WF

Ontologies

Ontology manager

Resource Discovery Services

UDDI/MDS

Metadata WSDL

Workflow Manager

- WF Translator
- WF Browser
- WF Scheduler
- WF Monitor

Grid Middleware

WS_1 WS_2 Spectra Management Services

WS_1 WS_2 Spectra Preprocessing Services

WS_1 WS_2 Data Mining Services

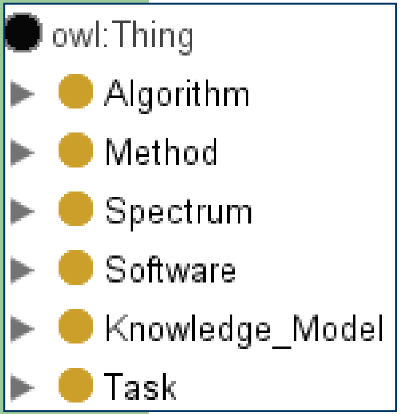
WS_1 WS_2 Spectra Preparation Services

WS_1 WS_2 Spectra Visualization Services

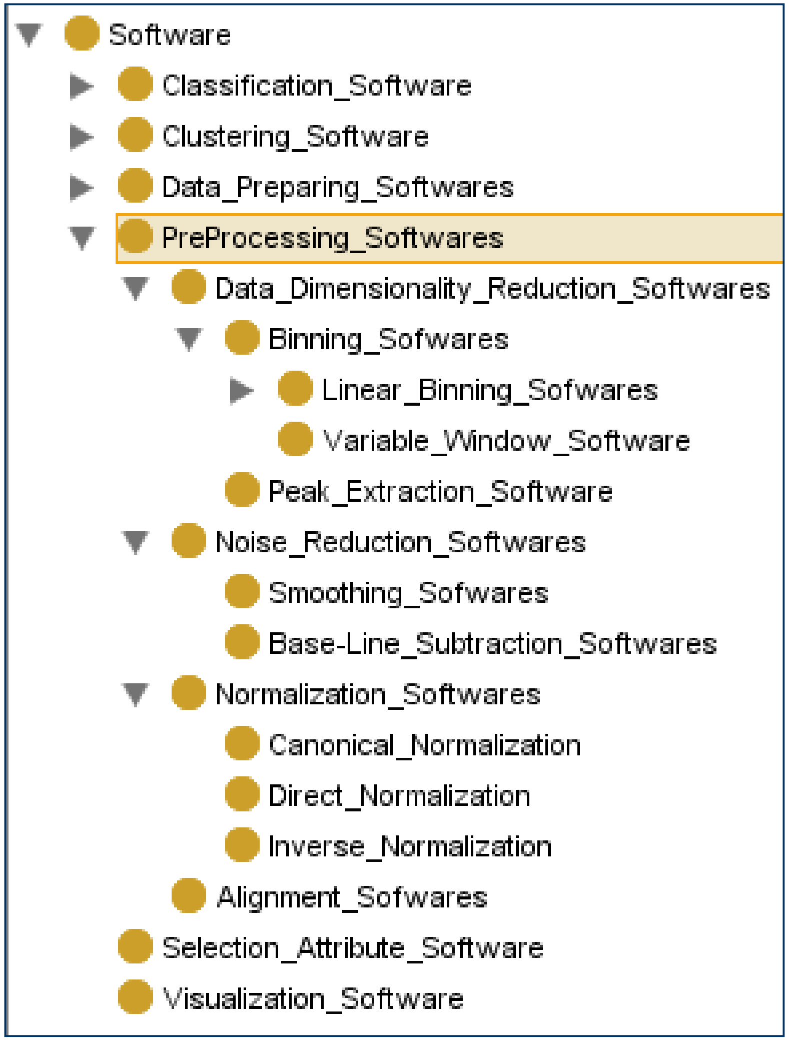
SpecDB APIs

- RSR: raw MS spectra
- PSR: pre-processed MS spectra
- PPSR: prepared pre-processed MS spectra

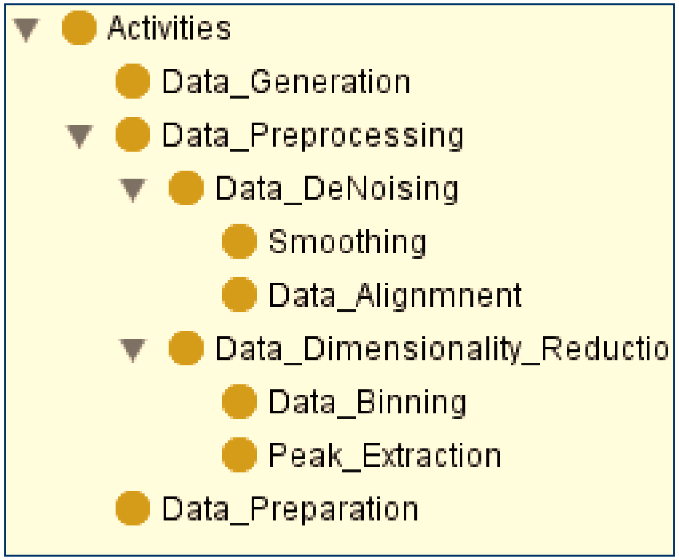
MS-Analyzer Ontologies



- owl:Thing
  - Algorithm
  - Method
  - Spectrum
  - Software
  - Knowledge_Model
  - Task

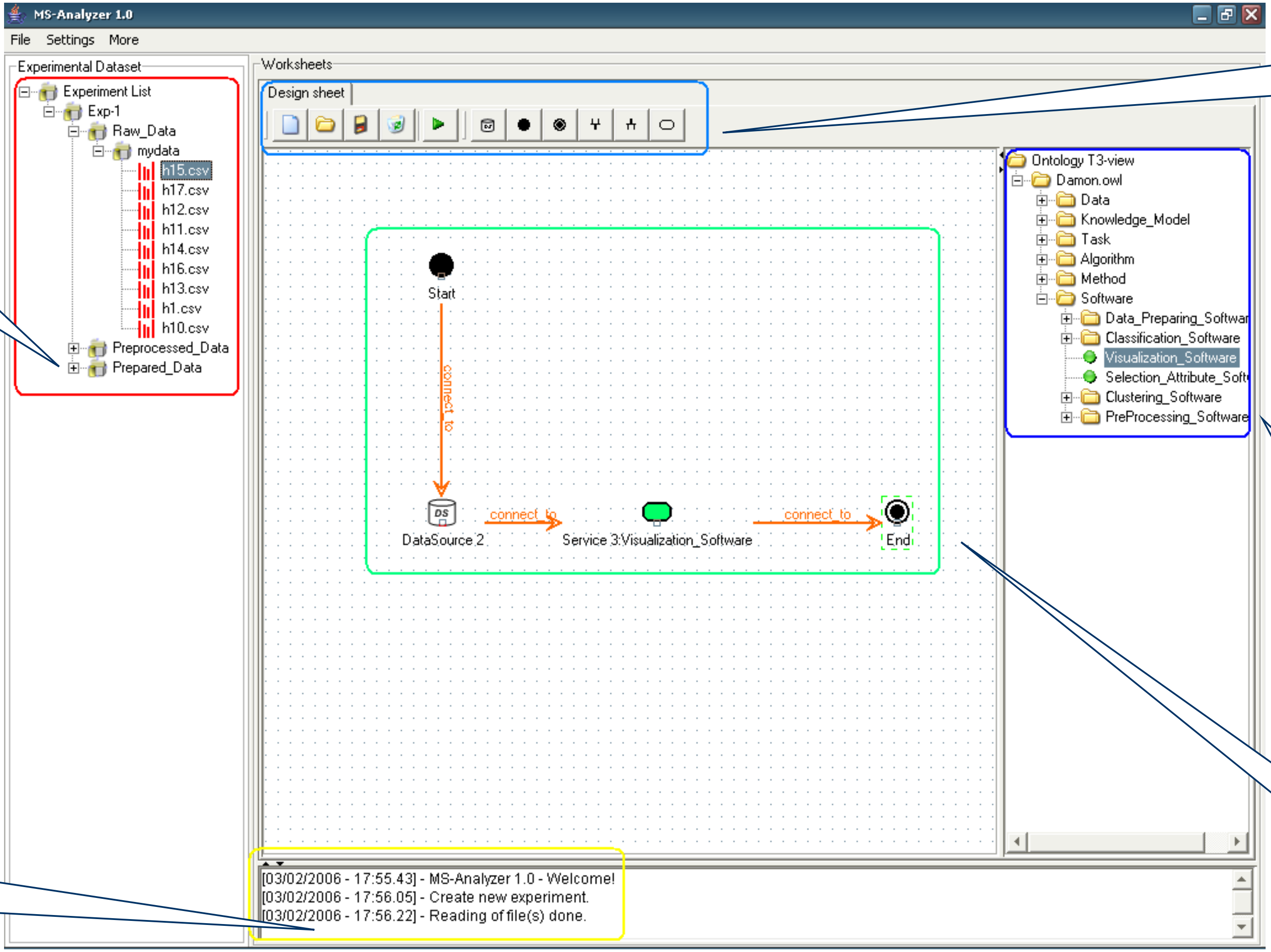


- Software
  - Classification_Software
  - Clustering_Software
  - Data_Preparing_Softwares
  - PreProcessing_Softwares
    - Data_Dimensionality_Reduction_Softwares
      - Binning_Sofwares
        - Linear_Binning_Sofwares
        - Variable_Window_Software
      - Peak_Extraction_Software
    - Noise_Reduction_Softwares
      - Smoothing_Sofwares
      - Base-Line_Subtraction_Softwares
    - Normalization_Softwares
      - Canonical_Normalization
      - Direct_Normalization
      - Inverse_Normalization
    - Alignment_Sofwares
  - Selection_Attribute_Software
  - Visualization_Software



- Activities
  - Data_Generation
  - Data_Preprocessing
    - Data_DeNoising
      - Smoothing
      - Data_Alignmnent
    - Data_Dimensionality_Reductio
      - Data_Binning
      - Peak_Extraction
  - Data_Preparation

MS-Analyzer GUI



```
[03/02/2006 - 17:55.43] - MS-Analyzer 1.0 - Welcome!
[03/02/2006 - 17:56.05] - Create new experiment.
[03/02/2006 - 17:56.22] - Reading of file(s) done.
```

Design Sheet

Dataset Area

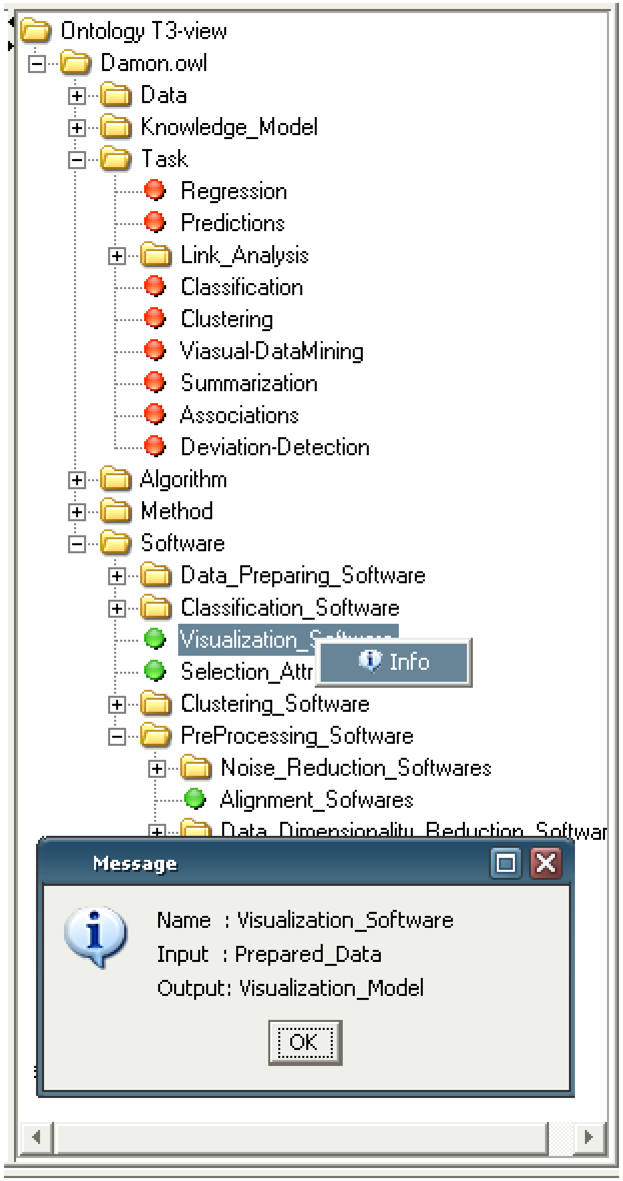
Ontology Area

Log Area

View Area

Ontology Manager

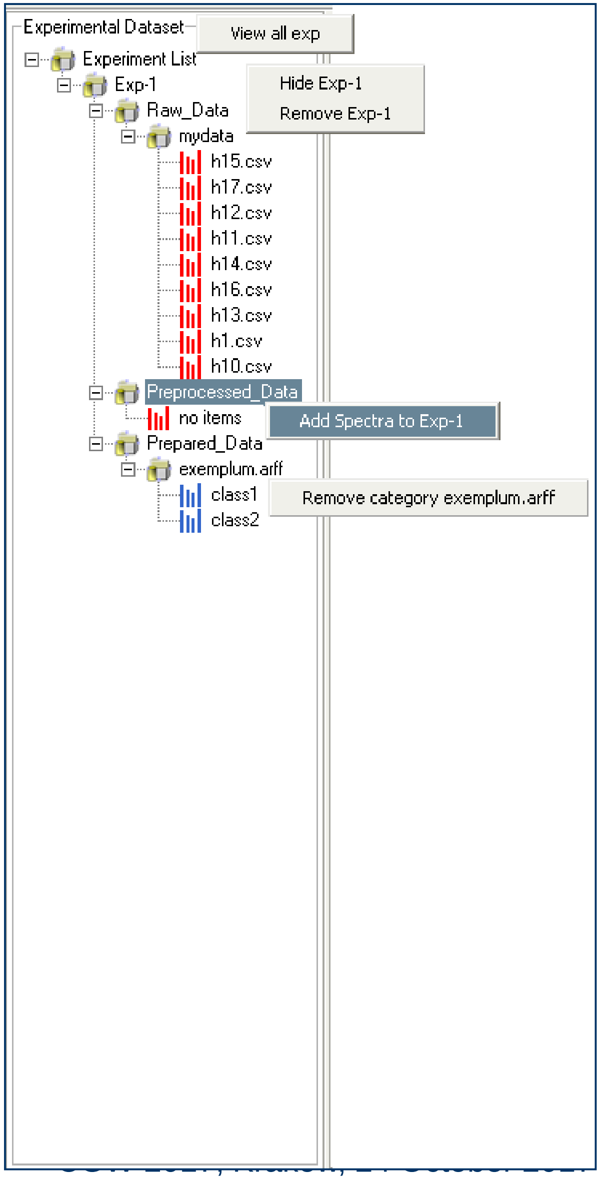
- Allows
  - loading user-defined OWL ontologies
  - browsing and searching bioinformatics tools
  - green services can be used in the editor by drag&drop



```
Name  : Visualization_Software
Input  : Prepared_Data
Output: Visualization_Model
```

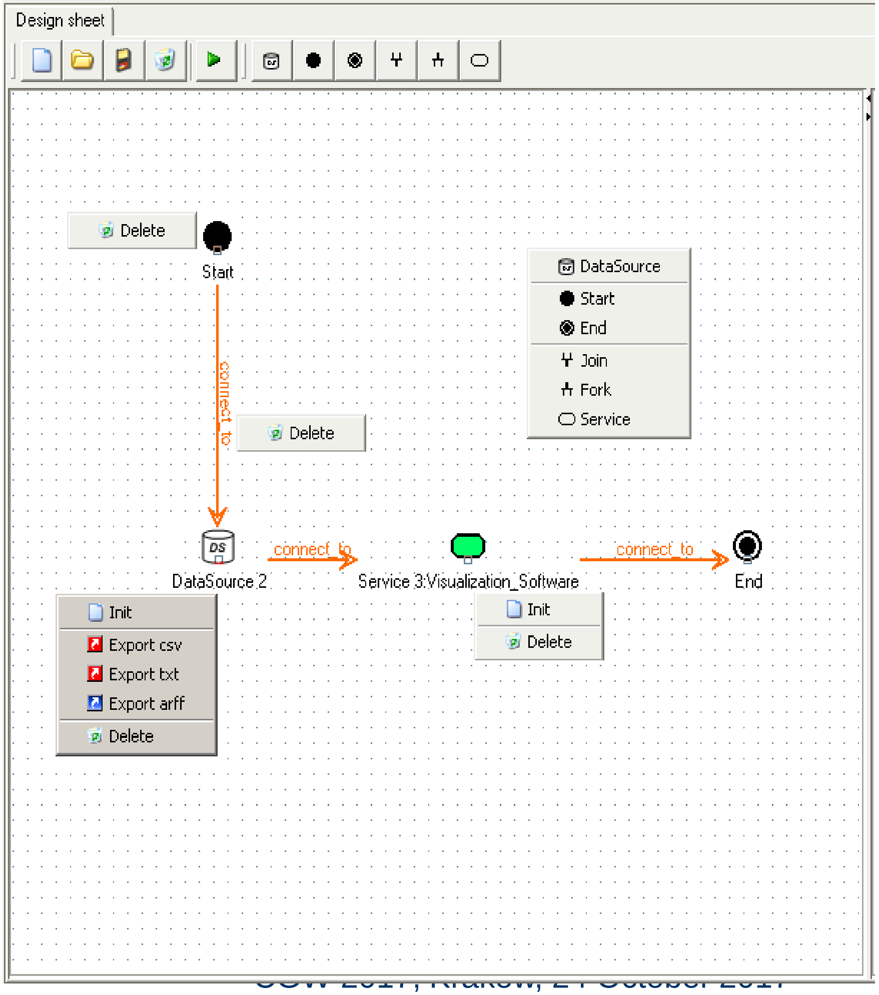
Dataset manager

- Each experiment is represented as a tree containing
  - raw and preprocessed datasets
- Different source spectra are supported
  - peak list, TXT, CSV, ARFF
- Different formats to be given in input to data mining
  - ARFF, Clementine
- The user can
  - Add a new experiment
  - Loading of experiment



Workflow Editor

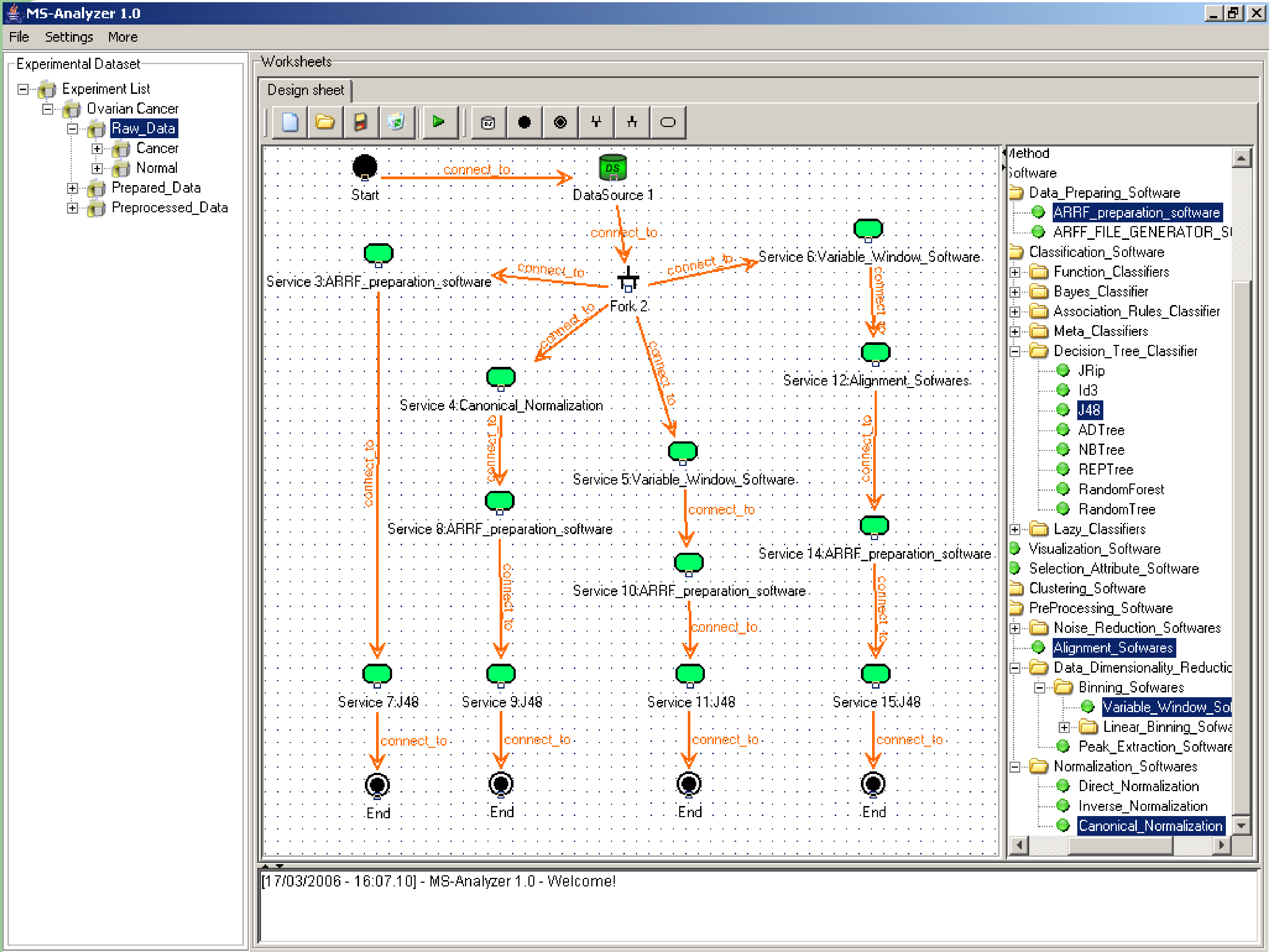
- UML-based notation
- Services are taken from the ontology pane
- Datasets are taken from the Dataset pane
- Constraints expressed by the ontology are enforced at composition time
- Ongoing: generation of a BPEL WF schema to be scheduled by a (Grid) workflow engine



U M G

Dubium sapientiae initium

M. Cannataro, P. H. Guzzi, T. Mazza, G. Tradigo, P. Veltri: Using ontologies for preprocessing and mining spectra data on the Grid. **Future Generation Comp. Syst.** 23(1): 55-60 (2007)

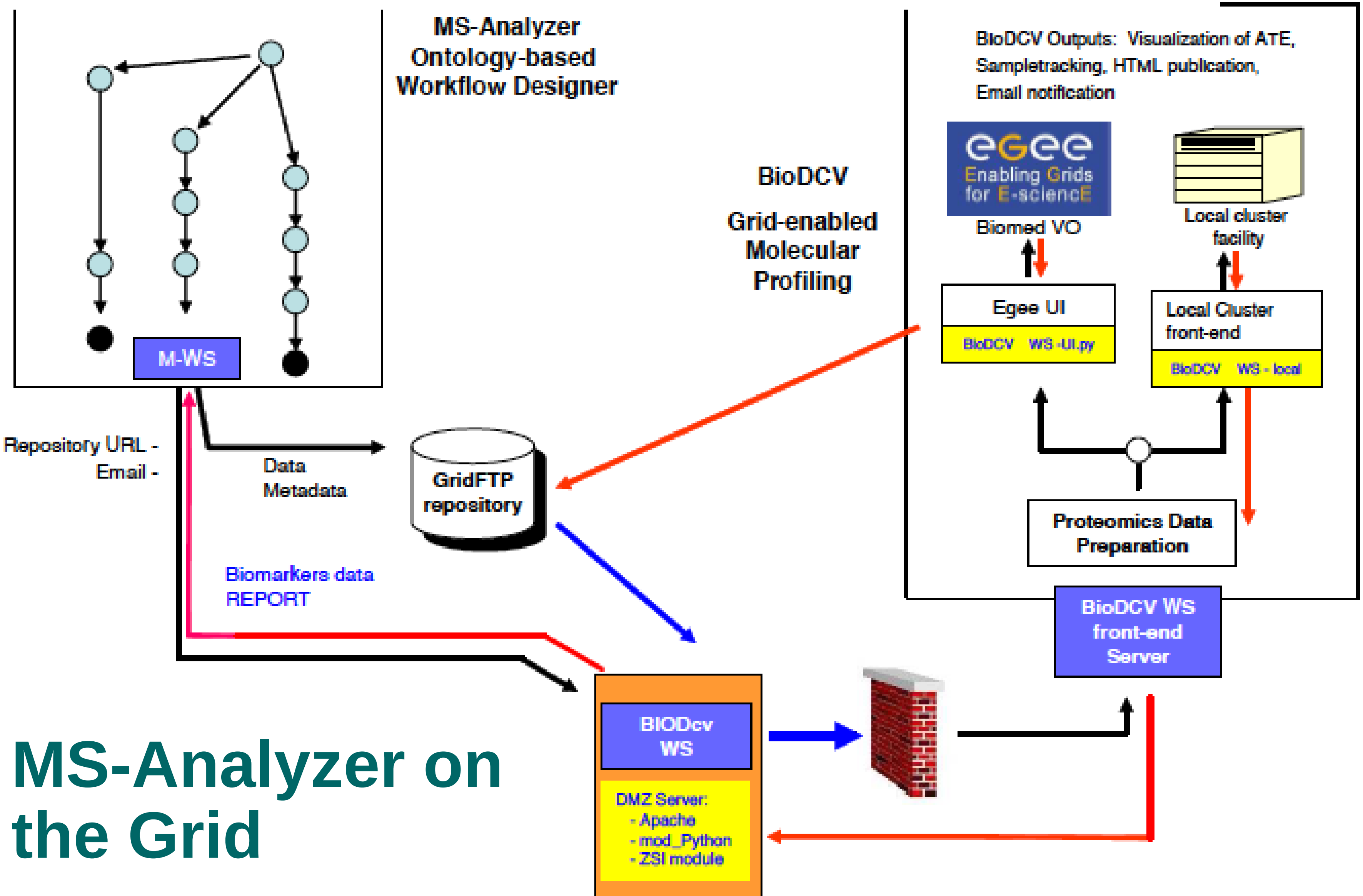


[17/03/2006 - 16:07.10] - MS-Analyzer 1.0 - Welcome!

IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 6, NO. 2, JUNE 2007 117

A Grid Environment for High-Throughput Proteomics

Mario Cannataro*, *Associate Member, IEEE*, Annalisa Barla, Roberto Flor, Giuseppe Jurman, Stefano Merler,
Silvano Paoli, Giuseppe Tradigo, Pierangelo Veltri, and Cesare Furlanello, *Associate Member, IEEE*

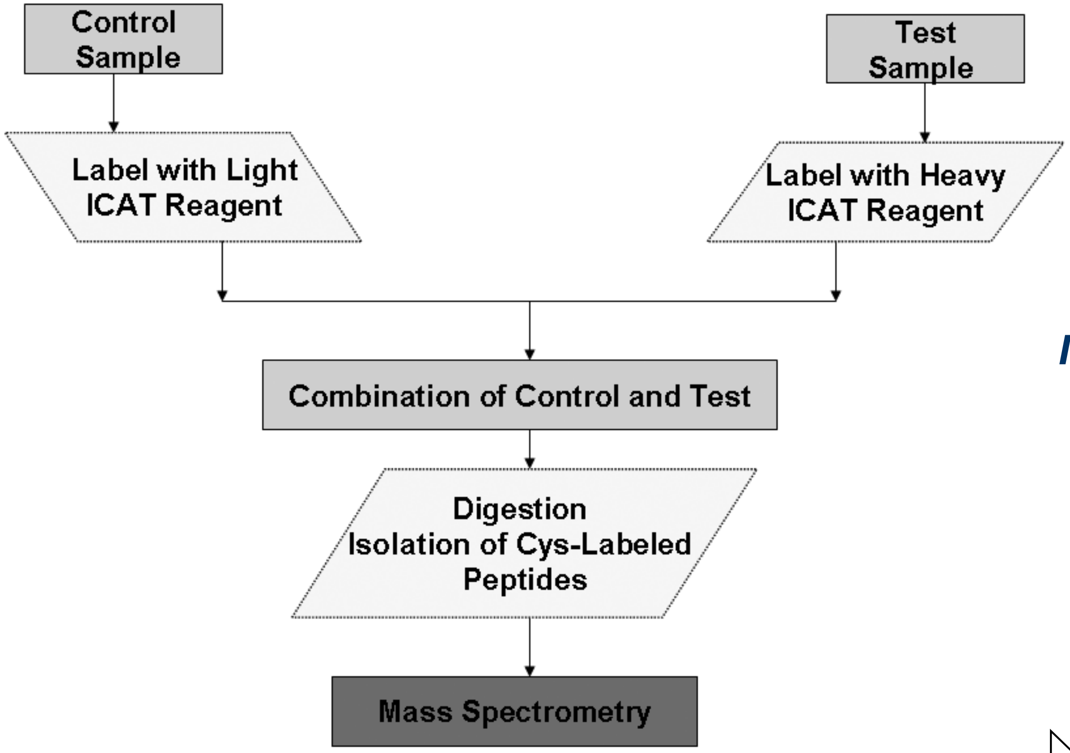


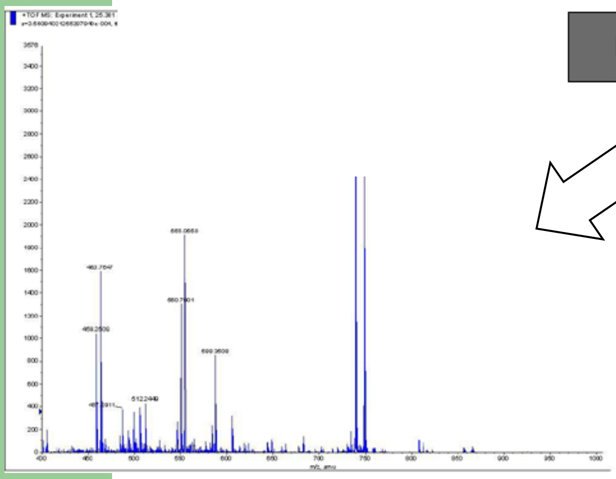
MS-Analyzer on the Grid

Protein Identification with Tandem Mass Spectrometry (MS/MS)

- Detection: first MS detects and selects the most abundant peptides (detection)
- Discovery: second MS analyzes such peptides producing a new spectrum
- Identification/Quantitation: the list of peaks is used to query different protein/peptide identification tools
  - MASCOT, ProteinProspector, Sequest
- To obtain a **quantitative** measurement about the protein abundance, MS may be coupled with opportune sample preparation protocols.
  - e.g. proteins can be identified by using tandem mass spectrometry with ICAT (Isotope Coded Affinity Tag) protocol for sample preparation.

ICAT + MS/MS protein identification


Control Sample
Test Sample
Label with Light ICAT Reagent
Label with Heavy ICAT Reagent
Combination of Control and Test
Digestion
Isolation of Cys-Labeled Peptides
Mass Spectrometry

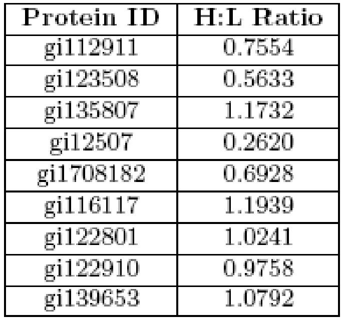


Pro ICAT

Swiss Prot DB

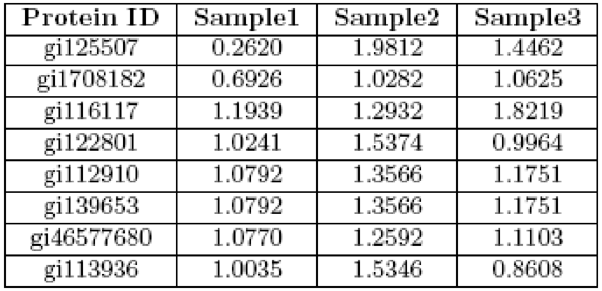
Identified and Quantified Proteins

Table 2. Single Sample Dataset



Protein ID	H:L Ratio
gi112911	0.7554
gi123508	0.5633
gi135807	1.1732
gi12507	0.2620
gi1708182	0.6928
gi116117	1.1939
gi122801	1.0241
gi122910	0.9758
gi139653	1.0792

Table 3. Multiple Samples Dataset

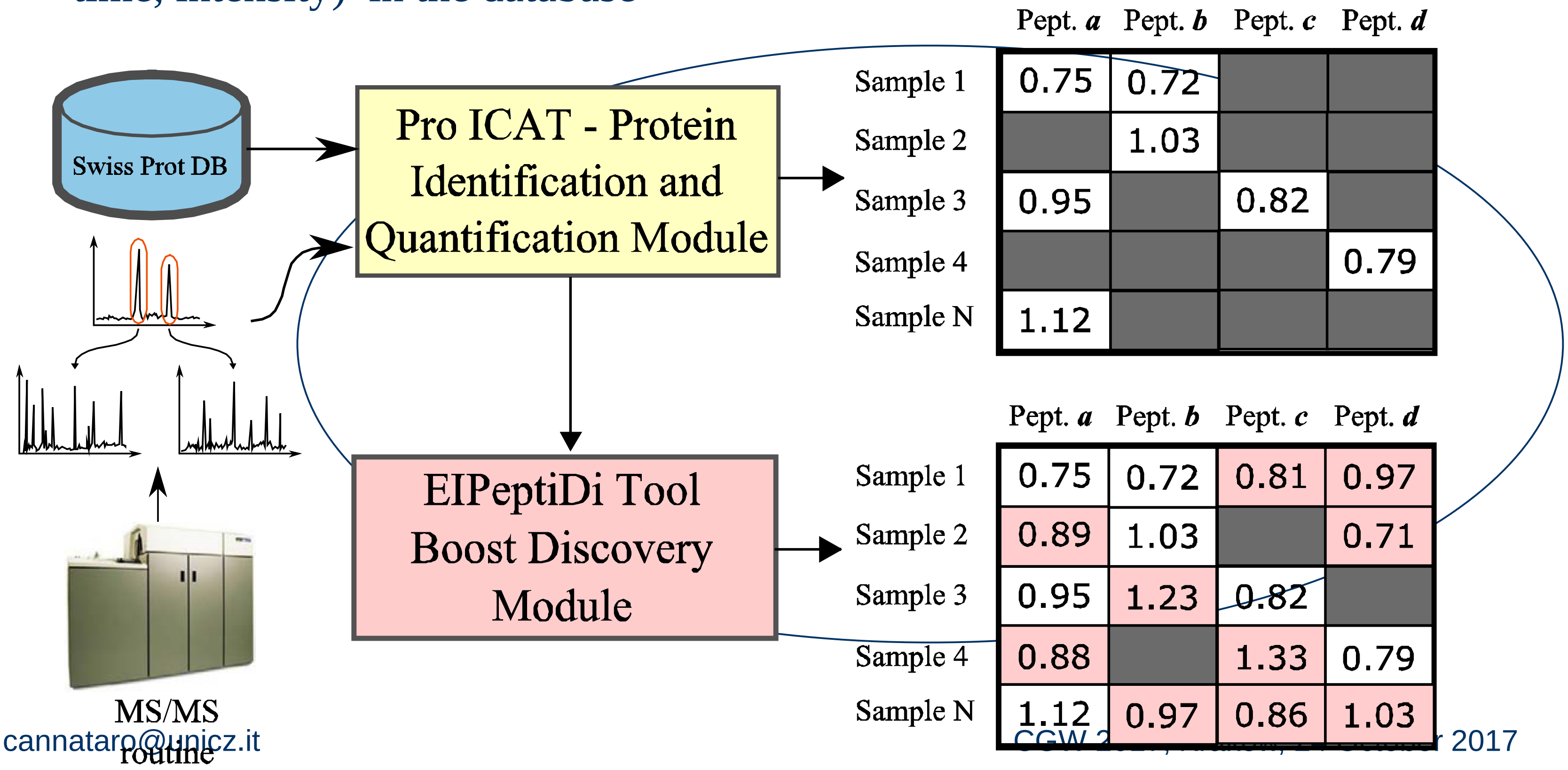


Protein ID	Sample1	Sample2	Sample3
gi125507	0.2620	1.9812	1.4462
gi1708182	0.6926	1.0282	1.0625
gi116117	1.1939	1.2932	1.8219
gi122801	1.0241	1.5374	0.9964
gi112910	1.0792	1.3566	1.1751
gi139653	1.0792	1.3566	1.1751
gi46577680	1.0770	1.2592	1.1103
gi113936	1.0035	1.5346	0.8608

Cannataro M, Cuda G, Gaspari M, Greco S, Tradigo G, Veltri P. The EIPeptiDi tool: enhancing peptide discovery in ICAT-based LC MS/MS experiments. **BMC Bioinformatics**. 2007 Jul 15;8:255.

EiPeptiDi: Enhanced ICAT Peptide Identification Discovery Tool

- Identified peaks (m/Z, retention time, intensity) + name of originating peptide/protein are stored in a database
- In new experiments, **peaks detected by the first MS, but not selected for second MS** can be identified by searching for “similar” peaks (m/Z, retention time, intensity) in the database


Swiss Prot DB
Pro ICAT - Protein Identification and Quantification Module
EIPeptiDi Tool Boost Discovery Module
MS/MS routine

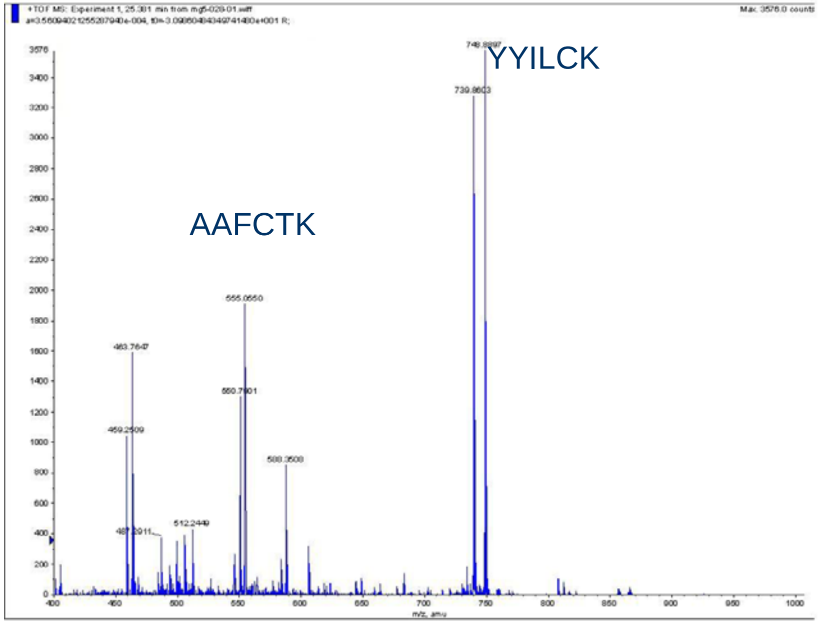
	Pept. *a*	Pept. *b*	Pept. *c*	Pept. *d*
Sample 1	0.75	0.72		
Sample 2		1.03		
Sample 3	0.95		0.82	
Sample 4				0.79
Sample N	1.12			

	Pept. *a*	Pept. *b*	Pept. *c*	Pept. *d*
Sample 1	0.75	0.72	0.81	0.97
Sample 2	0.89	1.03		0.71
Sample 3	0.95	1.23	0.82	
Sample 4	0.88		1.33	0.79
Sample N	1.12	0.97	0.86	1.03

EiPeptiDi Algorithm

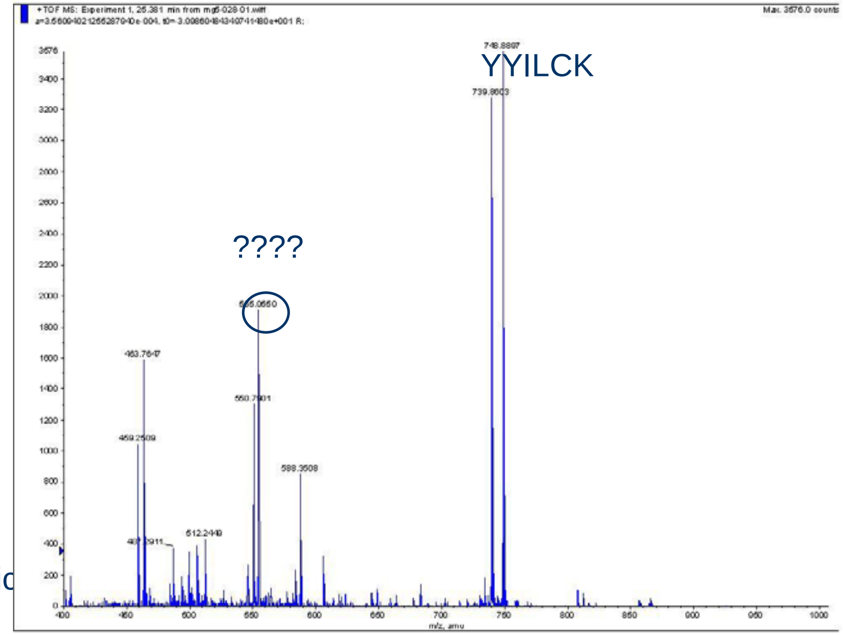
Sample S1:

2 peptides identified and quantified



Sample S2:

2 peptides quantified, but **only one identified**



EIPEPTIDI cross validates peptides identification across the data set

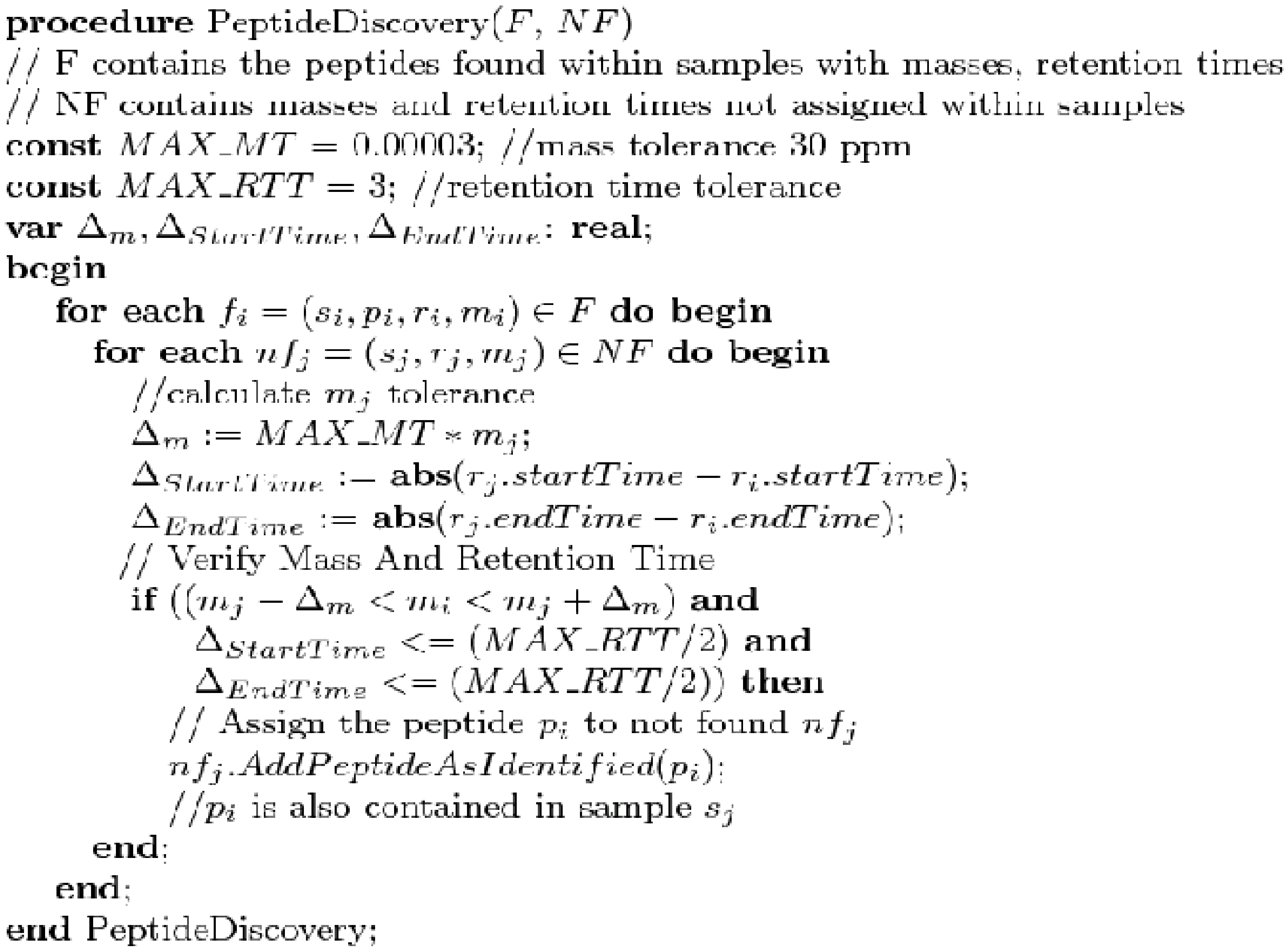
Retention time windows

$$St_i+delta<=rt_x<= St_i-delta$$

Mass windows

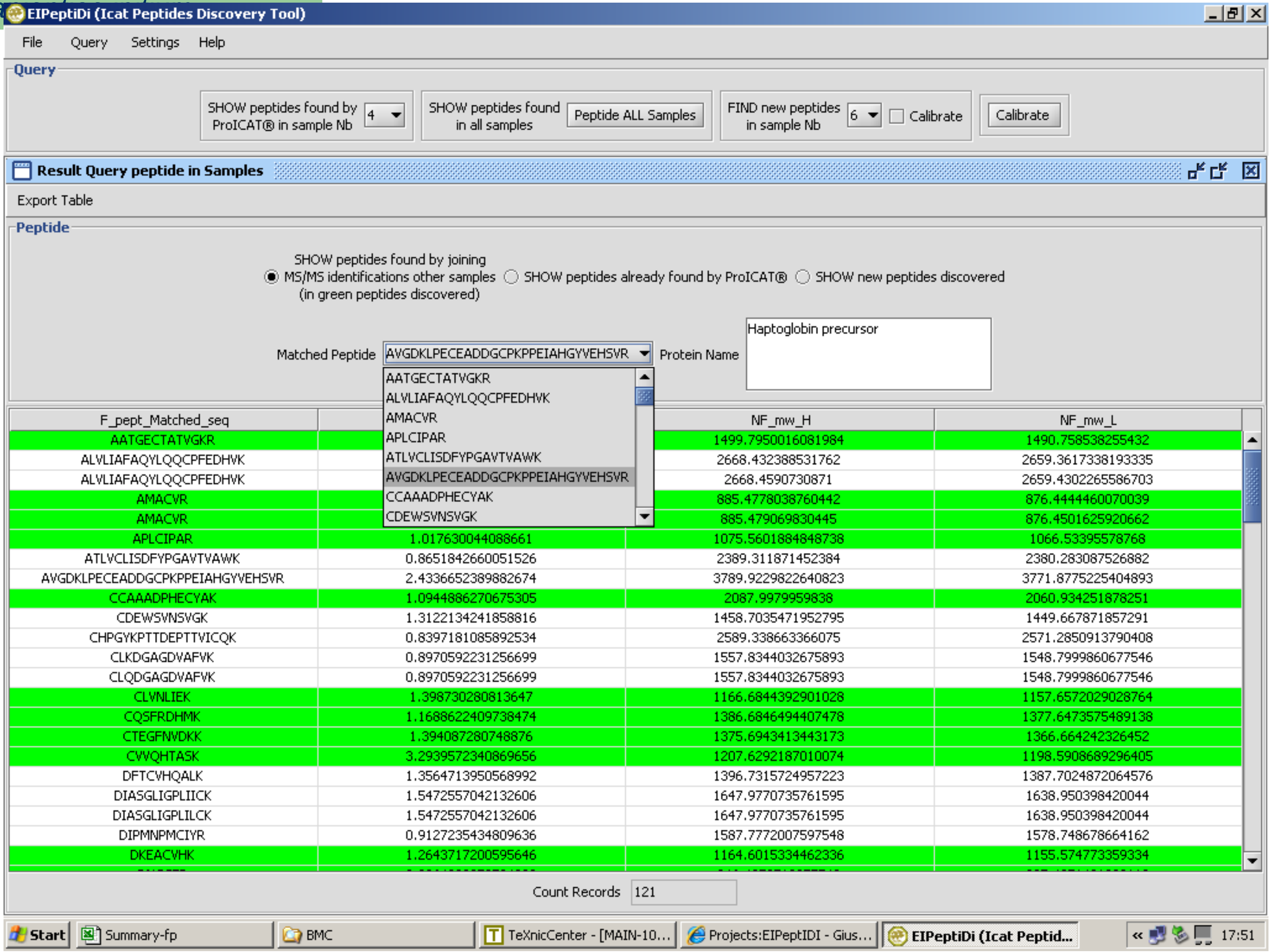
$$(Sm_i -(m_x*deltaMz)) <=MwF<= (Sm_i +(mx*deltaMz))$$

Algorithm



```
procedure PeptideDiscovery($F$, $NF$)
// F contains the peptides found within samples with masses, retention times
// NF contains masses and retention times not assigned within samples
const $MAX\_MT = 0.00003$; //mass tolerance 30 ppm
const $MAX\_RTT = 3$; //retention time tolerance
var $\Delta_m, \Delta_{StartTime}, \Delta_{EndTime}$: real;
begin
    for each $f_i = (s_i, p_i, r_i, m_i) \in F$ do begin
        for each $nf_j = (s_j, r_j, m_j) \in NF$ do begin
            //calculate $m_j$ tolerance
            $\Delta_m := MAX\_MT * m_j$;
            $\Delta_{StartTime} :=$ abs($r_j.startTime - r_i.startTime$);
            $\Delta_{EndTime} :=$ abs($r_j.endTime - r_i.endTime$);
            // Verify Mass And Retention Time
            if (($m_j - \Delta_m < m_i < m_j + \Delta_m$) and
                $\Delta_{StartTime} <= (MAX\_RTT/2)$ and
                $\Delta_{EndTime} <= (MAX\_RTT/2)$) then
              // Assign the peptide $p_i$ to not found $nf_j$
              $nf_j.AddPeptideAsIdentified(p_i)$;
              //$p_i$ is also contained in sample $s_j$
        end;
    end;
end PeptideDiscovery;
```

EiPeptiDi GUI



EIPeptiDi (Icat Peptides Discovery Tool)

Query

SHOW peptides found by ProICAT® in sample Nb: 4

SHOW peptides found in all samples: Peptide ALL Samples

FIND new peptides in sample Nb: 6 □ Calibrate

Result Query peptide in Samples

Peptide

● SHOW peptides found by joining MS/MS identifications other samples (in green peptides discovered) ○ SHOW peptides already found by ProICAT® ○ SHOW new peptides discovered

Matched Peptide: AVGDKLPECEADDGCPKPPEIAHGYVEHSVR

Protein Name: Haptoglobin precursor

AATGECTATVGKR
ALVLIAFAQYLQQCPFEDHVK
AMACVR
APLCIPAR
ATLVCLISDFYPGAVTVAWK
AVGDKLPECEADDGCPKPPEIAHGYVEHSVR
CCAAADPHECYAK
CDEWSVNSVGK

F_pept_Matched_seq		NF_mw_H	NF_mw_L
AATGECTATVGKR		1499.7950016081984	1490.758538255432
ALVLIAFAQYLQQCPFEDHVK		2668.432388531762	2659.3617338193335
ALVLIAFAQYLQQCPFEDHVK		2668.4590730871	2659.4302265586703
AMACVR		885.4778038760442	876.4444460070039
AMACVR		885.479069830445	876.4501625920662
APLCIPAR	1.017630044088661	1075.5601884848738	1066.53395578768
ATLVCLISDFYPGAVTVAWK	0.8651842660051526	2389.311871452384	2380.283087526882
AVGDKLPECEADDGCPKPPEIAHGYVEHSVR	2.4336652389882674	3789.9229822640823	3771.8775225404893
CCAAADPHECYAK	1.0944886270675305	2087.9979959838	2060.934251878251
CDEWSVNSVGK	1.3122134241858816	1458.7035471952795	1449.667871857291
CHPGYKPTTDEPTTVICQK	0.8397181085892534	2589.338663366075	2571.2850913790408
CLKDGAGDVAFVK	0.8970592231256699	1557.8344032675893	1548.7999860677546
CLQDGAGDVAFVK	0.8970592231256699	1557.8344032675893	1548.7999860677546
CLVNLIEK	1.398730280813647	1166.6844392901028	1157.6572029028764
CQSFRDHMK	1.1688622409738474	1386.6846494407478	1377.6473575489138
CTEGFNVDKK	1.394087280748876	1375.6943413443173	1366.664242326452
CVVQHTASK	3.2939572340869656	1207.6292187010074	1198.5908689296405
DFTCVHQALK	1.3564713950568992	1396.7315724957223	1387.7024872064576
DIASGLIGPLIICK	1.5472557042132606	1647.9770735761595	1638.950398420044
DIASGLIGPLILCK	1.5472557042132606	1647.9770735761595	1638.950398420044
DIPMNPMCIYR	0.9127235434809636	1587.7772007597548	1578.748678664162
DKEACVHK	1.2643717200595646	1164.6015334462336	1155.574773359334

Count Records: 121

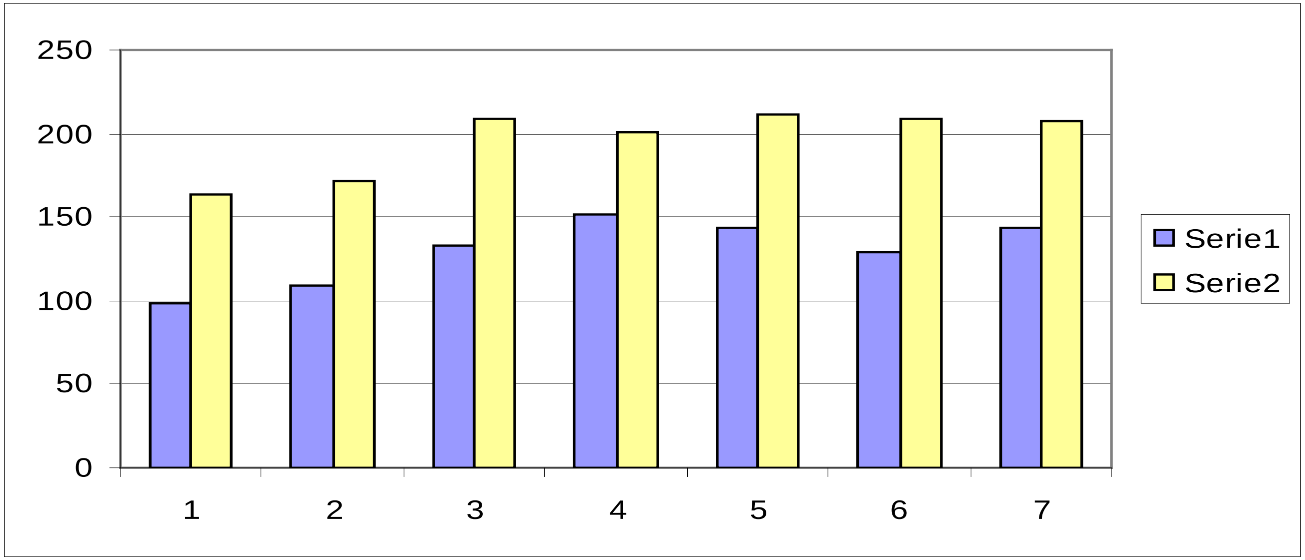
Start
Summary-fp
BMC
TeXnicCenter - [MAIN-10...
Projects:EIPeptIDI - Gius...
EIPeptiDi (Icat Peptid...
17:51

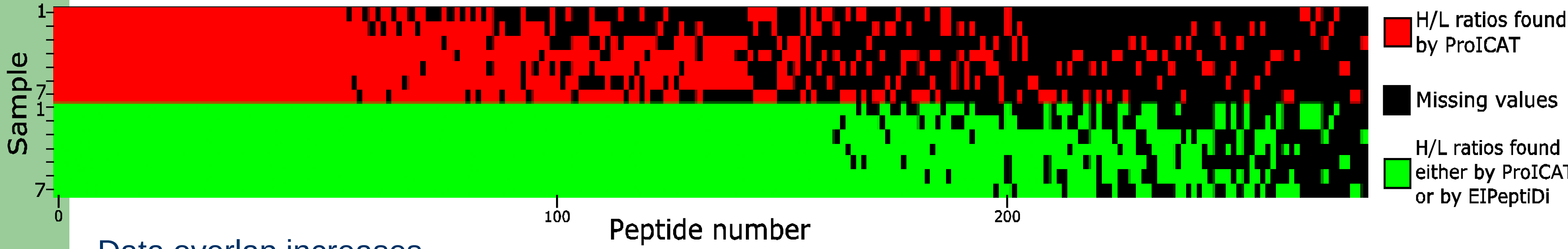
EiPeptiDi is available on line at:

http://bioingegneria.unicz.it/~veltri/projects/eipeptidi

Experimental Results

- In a 7 samples data sets we increase the number of Identified Peptides of 50 % (av) using EiPeptiDi.





Data overlap increases

Experimentation

- EIPETIDI has been successfully used to improve the identification of proteins used to discriminate between different classes of familial adenomatous polyposis (FAP) patients
  - Barbara Quaresima, Telma Crugliano, Marco Gaspari, Maria Concetta Faniello, Paola Cosimo, Rosa Valanzano, Maurizio Genuardi, Mario Cannataro, Pierangelo Veltri, Francesco Baudi, Patrizia Doldo, Giovanni Cuda, Salvatore Venuta, Francesco Costanzo, A proteomics approach to identify changes in protein profiles in serum of familial adenomatous polyposis patients, Cancer Letters 272(1):40-52,8 December 2008, http://dx.doi.org/10.1016/j.canlet.2008.06.021

OUTLINE

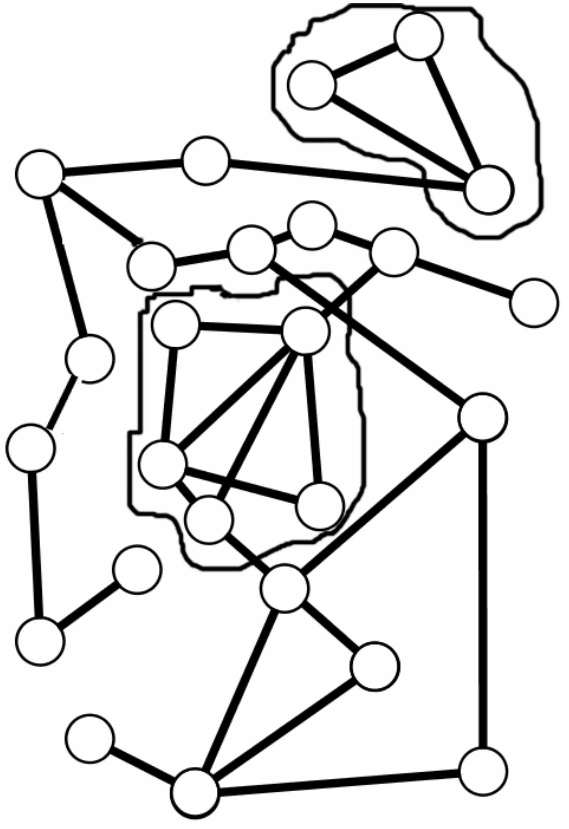
- Experiences at University of Catanzaro in high performance management, preprocessing and analysis of omics data
  - PART I: Genomics data
  - PART II: Proteomics data
  - **PART III: Interactomics data**
- Conclusions

Interactomics Data

- Interactomics
  - Protein-to-Protein Interactions (PPI)
  - PPI Databases and Standards
  - Protein Complex Prediction
- Experiences at University of Catanzaro:
  - **IMPRECO**: a meta-predictor for protein complexes
  - **CytoMCL**: Markov Clustering of Metabolic Networks in Cytoscape
  - **CytoSeVis**: Semantic Similarity-Based Visualization of PPI Networks in Cytoscape
  - **ONTOPIN**: using ontologies for querying PPI databases

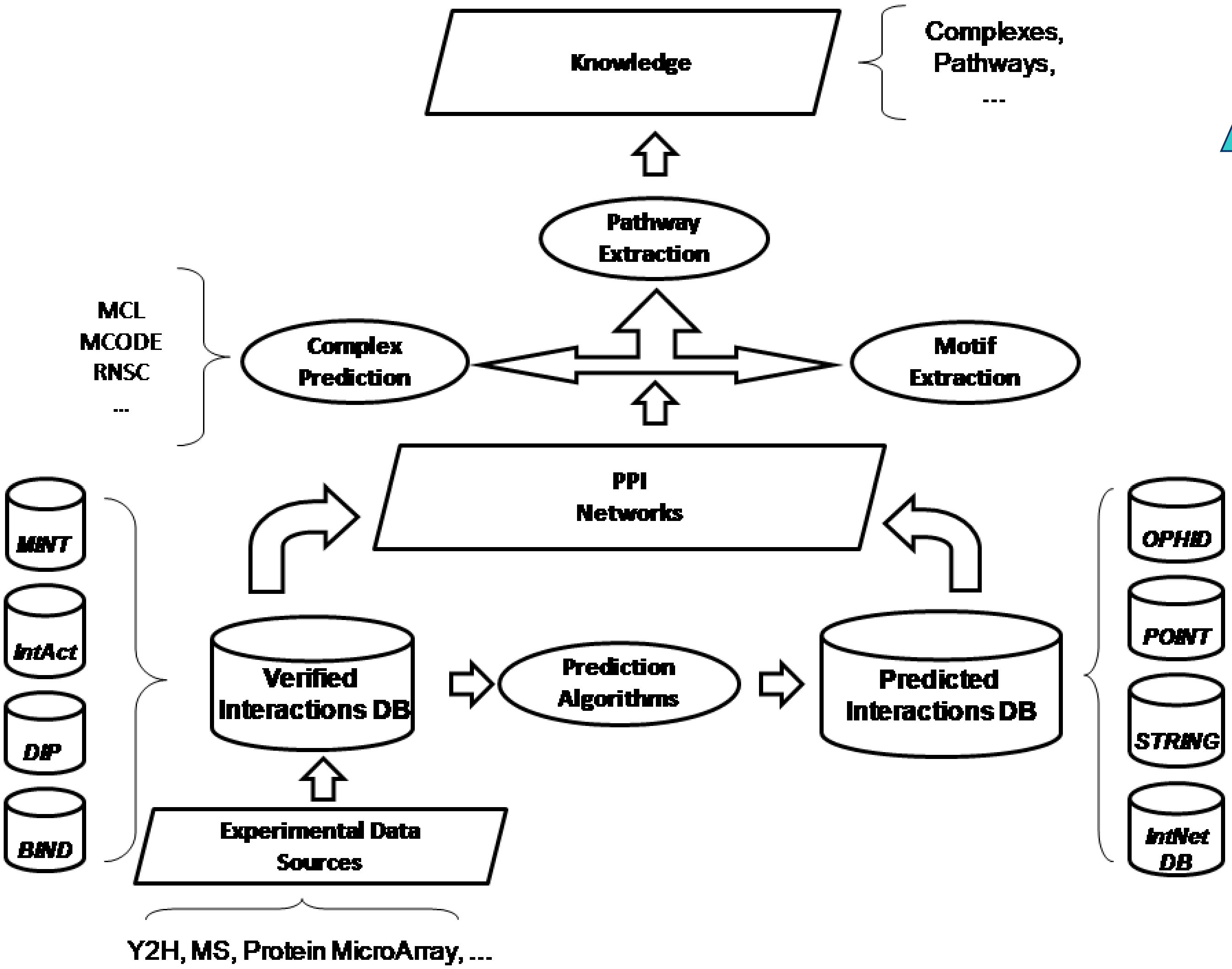
Interactomics

- **Interactomics** is the study of the whole set of protein interactions
  - Protein to Protein Interaction (PPI) network.
- PPI Networks are usually modeled as undirect graphs.
  - Global analysis investigated main properties.
- Different network models have been proposed:
  - Scale Free
  - Random Graphs
  - Geometric Random Graph



M. CANNATARO, P. H. GUZZI, and P. VELTRI, Protein-to-Protein Interactions: Technologies, Databases, and Algorithms, ACM Computing Surveys, 43(1), November 2010

Flow of Information in Interactomics



From wet lab to in silico analysis and prediction

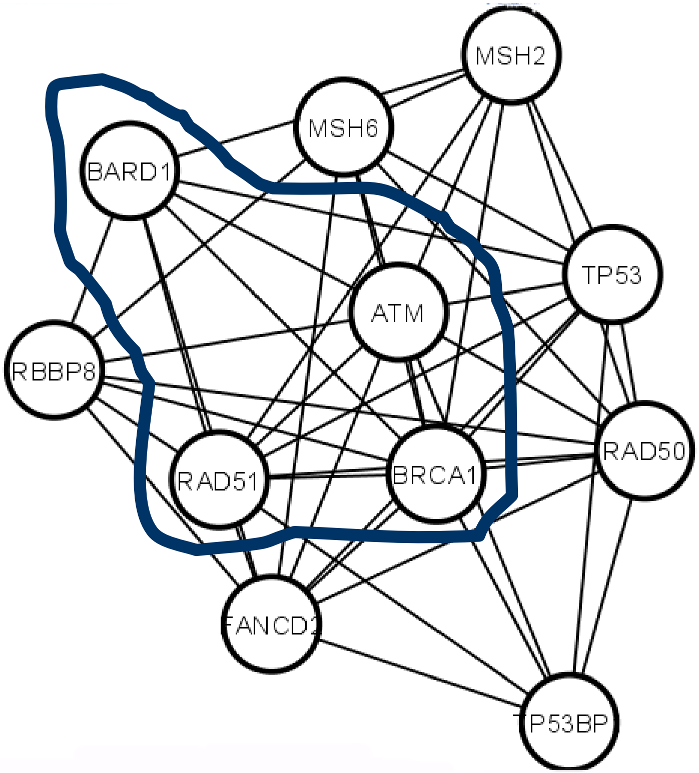
PPI Databases & Standards

- Verified Interactions DBs store experimentally determined (Yeast 2 Hybrid, Protein Micro Arrays, Mass Spectrometry) PPI data
  - **DIP** (Database of Interacting Proteins), **BIND** (Biomolecular Interaction Network Database), **MIPS**, **MPCDB, CORUM**
- Predicted Interactions DBs store interactions predicted in silico:
  - **OPHID** maps experimental interactions determined in model organisms into human interactions.
  - **POINT** projects verified interactions into human orthologs and filters interactions considering functional information.
  - **IntNetDB** predicts interaction by integrating different information (mRNA, co-expression, sequence similarity)
- Standards
  - BioPax, IMEx, PSI-MI, SBML, SIF

Sylvain Brohée1 and Jacques van Helden, Evaluation of clustering algorithms for protein-protein interaction networks BMC Bioinformatics. 2006; 7: 488.

Complex Prediction Algorithms

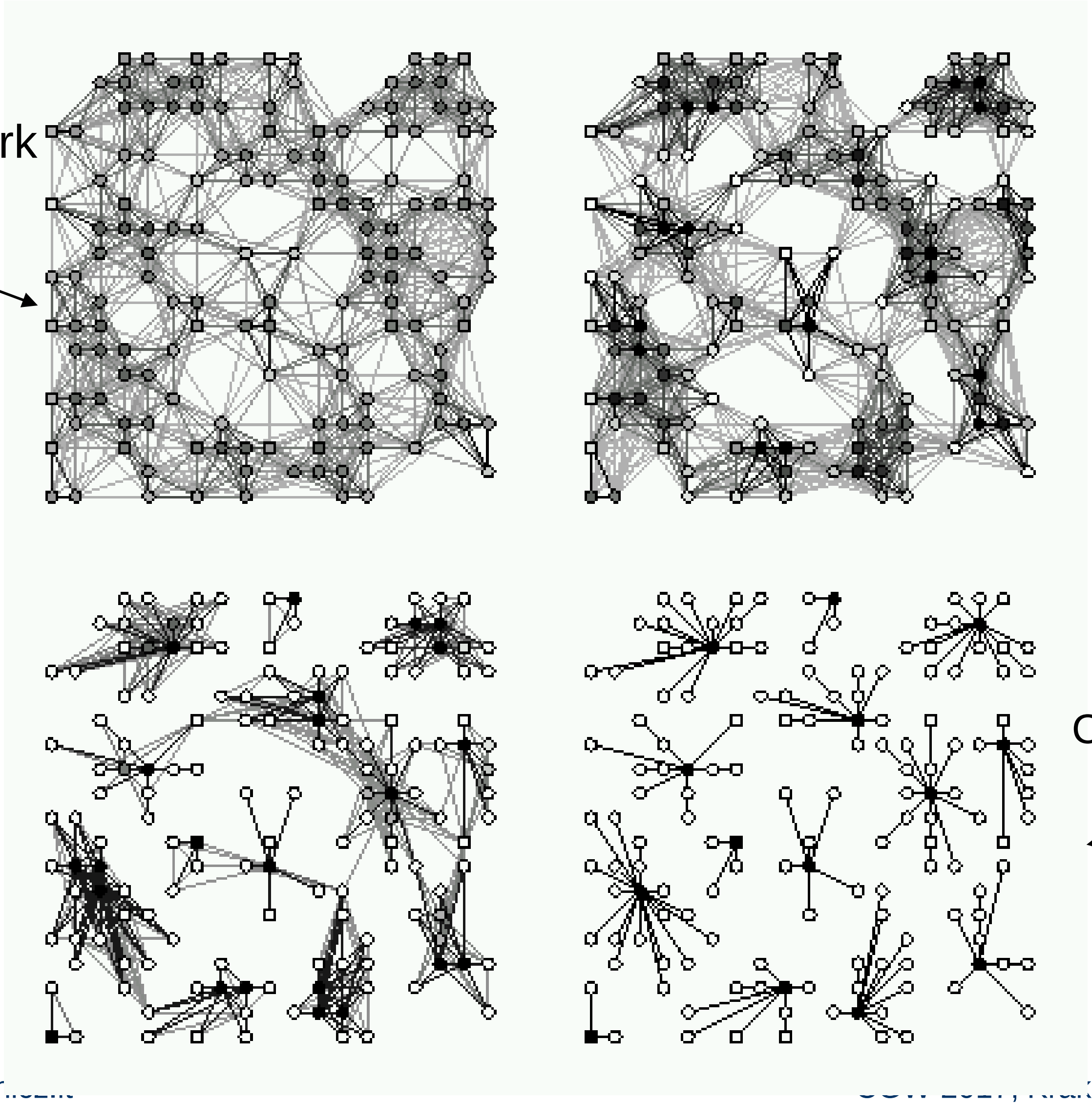
- A **protein complex** is a group of two or more associated proteins which interact sharing the same biological goal
  - i.e. a cluster in the PPI graph



- The Markov Cluster algorithm (MCL) [van Dongen 2000] finds clusters on a graph by simulating a stochastic flow and then analyzing its distribution
- The MCODE algorithm, takes in input an interaction network and tries to find complexes by building clusters.
- The Restricted Neighborhood Search (RNSC) predictor, after an initial random clustering, uses a cost-based local search algorithm based on the tabu heuristic

MCL

Input Network



Clustered Network

UMG

Dubium sapientiae initium

IMPRECO: A Tool for Improving the Prediction of Protein Complexes

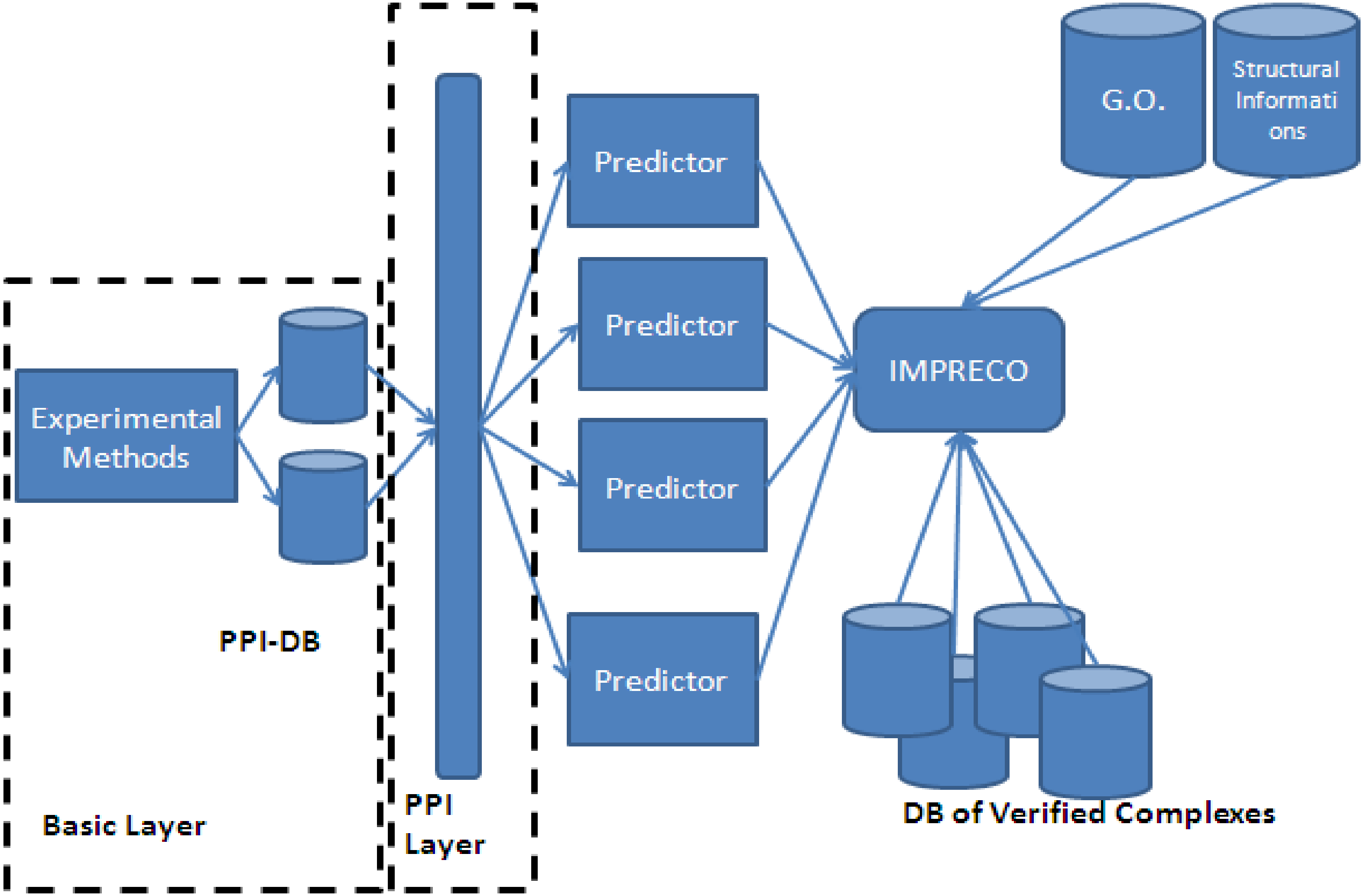
Mario Cannataro, Pietro H. Guzzi, Pierangelo Veltri

Bioinformatics Laboratory,

University "Magna Græcia" of Catanzaro, Italy

cannataro@unicz.it

IMPRECO: Improving the prediction of protein complexes

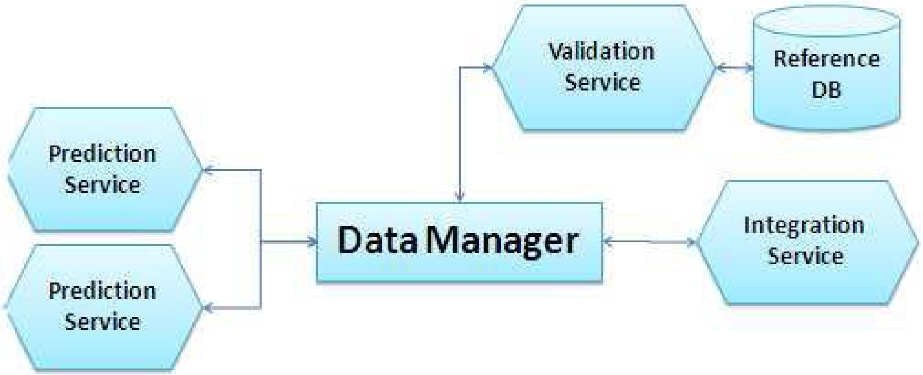


The IMPRECO Algorithm

- The rationale underlying the proposed meta-predictor is to combine different predictor results using an integration algorithm able to gather (partial) results from different predictors
- The integration algorithm starts by integrating results (i.e. clusters) obtained by running different available predictors.
- Three different cases are considered by evaluating the topological relations among clusters coming from the considered predictors:

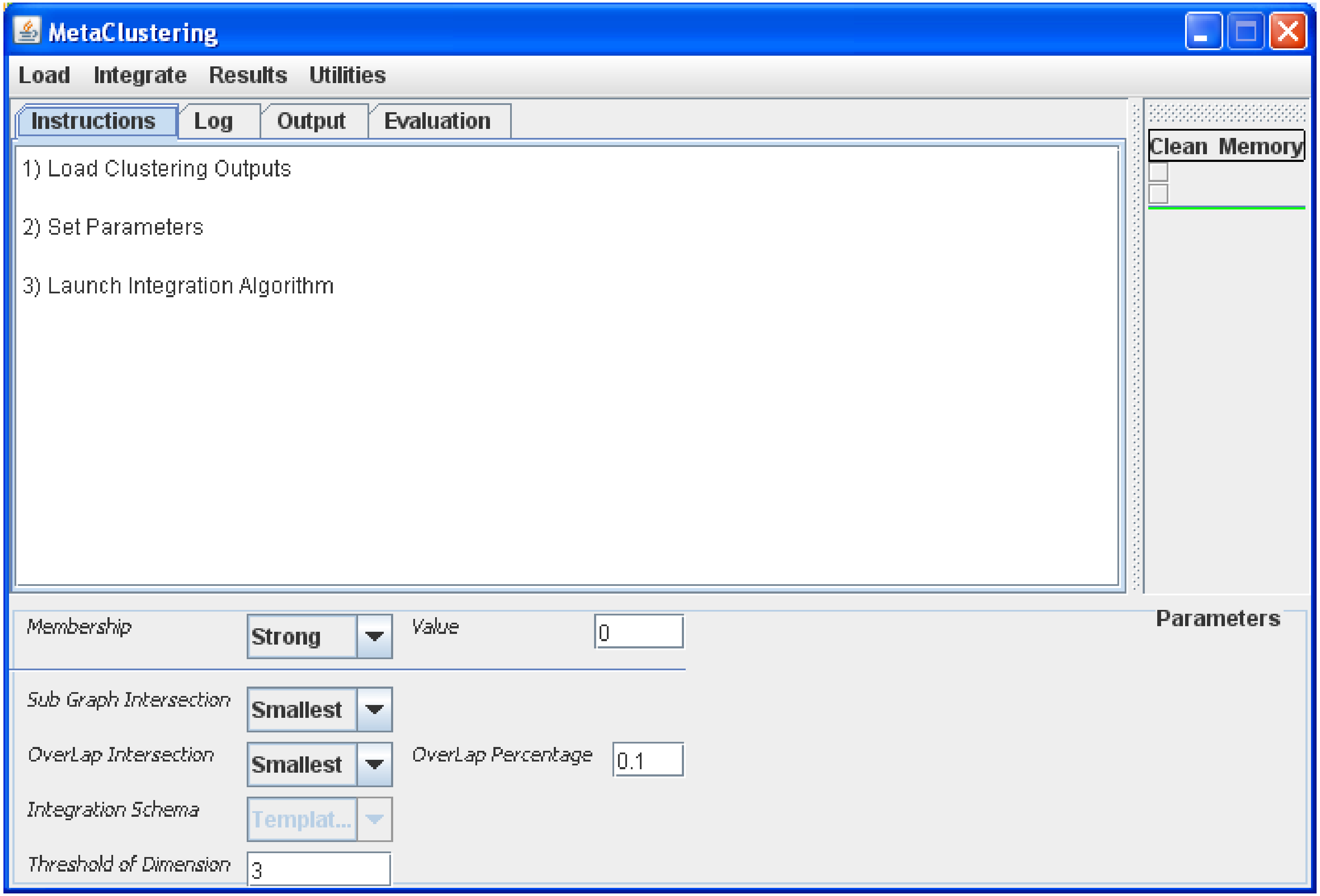
1. **equality:** the same clusters are returned by all (or by a significant number of) predictors,
2. **containment:** it is possible to identify a containment relation among (a set of) clusters returned by all (or by a significant number of) predictors;
3. **overlap:** it is possible to identify an overlap relation among (a set of) clusters returned by all (or by a significant number of) predictors;

Validation of Predicted Complexes



- To estimate the integration quality, IMPRECO uses an **evaluation module based on a reference database, i.e. a catalog of verified complexes** (e.g. the MIPS catalog)
- For each cluster the evaluation module calculates the measurements of sensitivity, positive predictive values and accuracy .
  - The first measure is an average representing the fraction of proteins of a complex that are found in a common cluster. When only a big cluster is found, the sensitivity tends to one.
  - The second measure represents the fraction of members of a cluster that belong to a given complex. When each protein belongs to one cluster, PPV is 1, conversely to the previous measure.
  - Thus, the third measure, being the geometric average of sensitivity and ppv represents a trade-off.

The IMPRECO TOOL



1) Load Clustering Outputs

2) Set Parameters

3) Launch Integration Algorithm

Membership: Strong

Value: 0

Sub Graph Intersection: Smallest

OverLap Intersection: Smallest

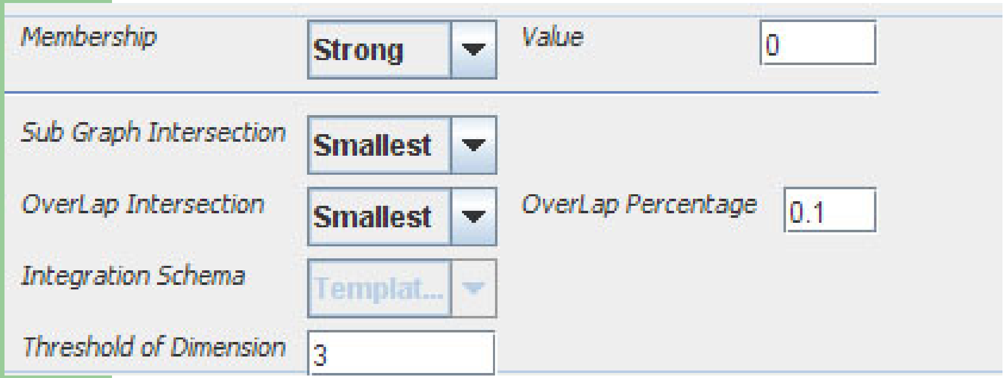
OverLap Percentage: 0.1

Integration Schema: Templat...

Threshold of Dimension: 3

Preliminary Results/1

In the first experiment we used a yeast network of 1094 nodes and 14658 edges and we run the MCL, RNSC and MCODE algorithms, obtaining respectively 165, 306 and 73 clusters.



Membership	Strong	Value	0
Sub Graph Intersection	Smallest		
OverLap Intersection	Smallest	OverLap Percentage	0.1
Integration Schema	Templat...		
Threshold of Dimension	3		

Parameter	Value
TM	1
IS	Biggest
II	Biggest
Overlap Percentage	0.1
TD	2

Parameter - Algorithm	MCODE	MCL	RNSC	IMPRECO
Clusters No.	73	165	306	155
Sensitivity	0,34	0,47	0,46	0,89
PPV	0,48	0,69	0,59	0,59
Accuracy	0,40	0,57	0,52	0,70

UMG

Dubium sapientiae initium

CytoMCL: Markov Clustering of Metabolic Networks in Cytoscape

Mario Cannataro, Pietro H. Guzzi

Bioinformatics Laboratory,

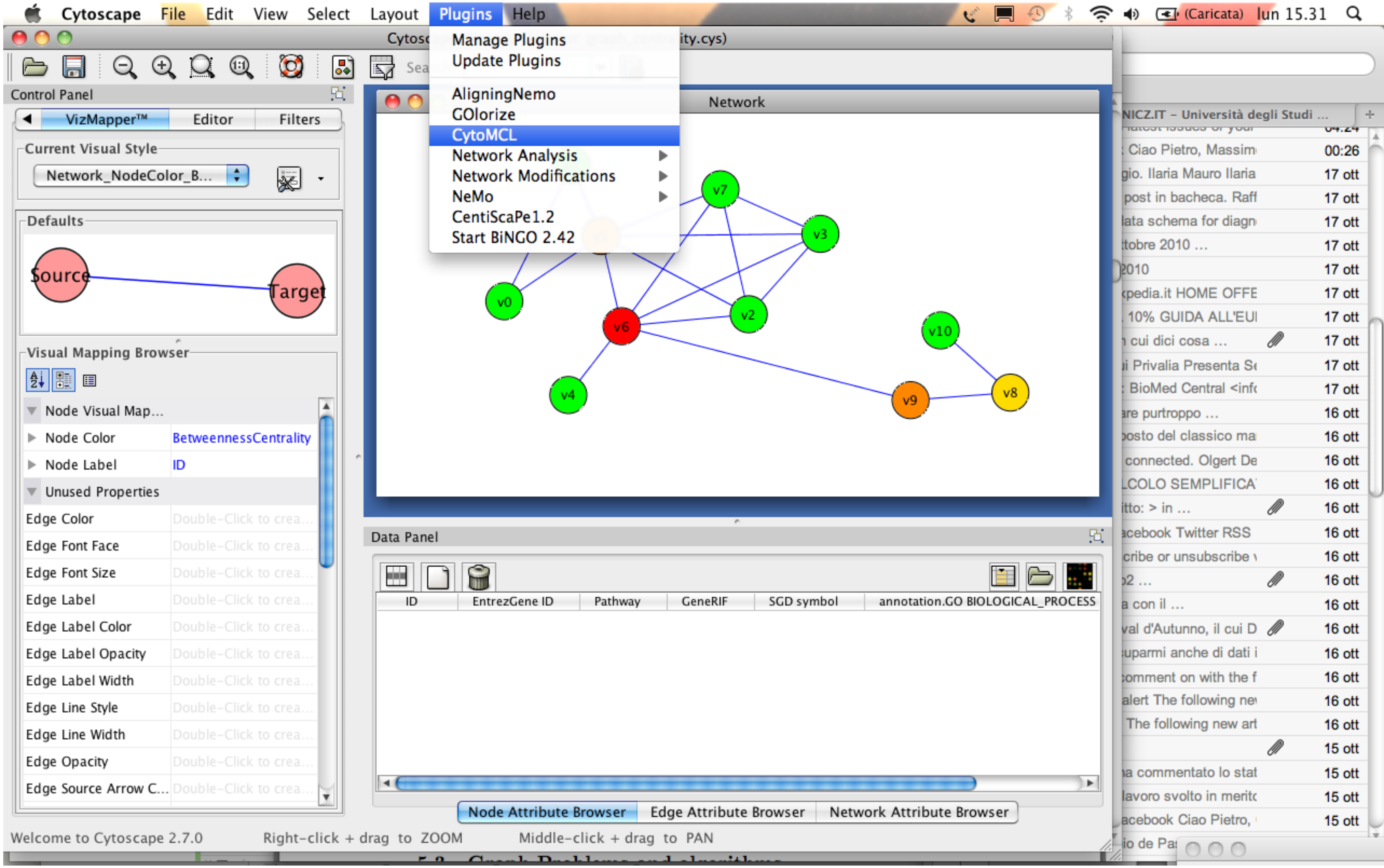
University “Magna Græcia” of Catanzaro, Italy

cannataro@unicz.it

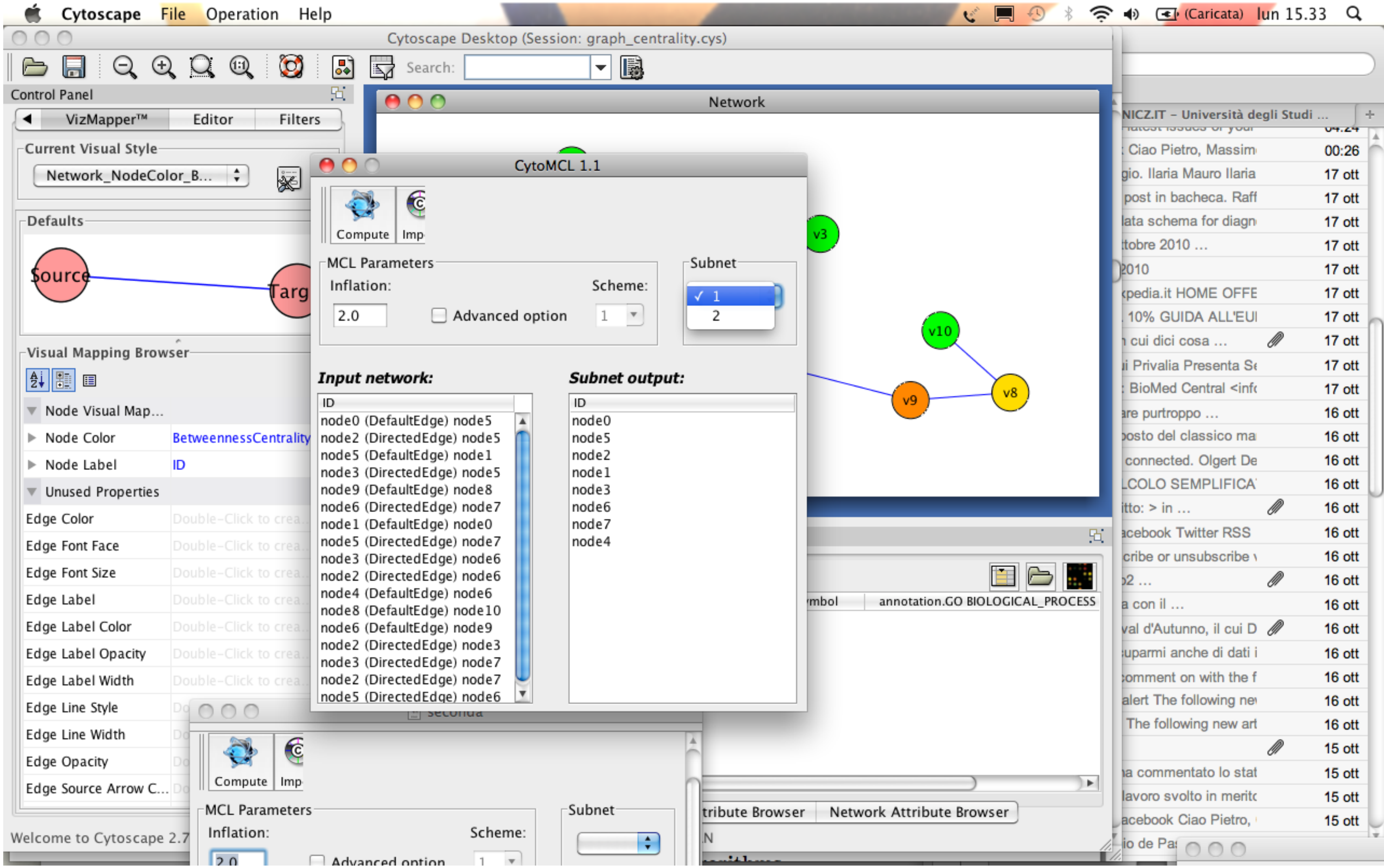
CytoMCL rationale

- The Markov Clustering Algorithm (MCL) is a well-known algorithm for clustering graphs.
- Cytoscape is a tool for visualizing and analyzing networks based on an extensible architecture.
- Nevertheless MCL does not provide a graphical user interface and cannot be used in the Cytoscape platform
- CytoMCL is a Cytoscape plugin that finds clusters in a graph by using the Markov Clustering Algorithm.
  - Based on an intuitive interface it is able to load a network from Cytoscape, to analyze it and to visualize resulting clusters into Cytoscape.

Starting the plugin



Running CytoMCL



CytoMCL 1.1

MCL Parameters

Inflation: 2.0

□ Advanced option

Scheme: 1

Subnet: ✓ 1 / 2

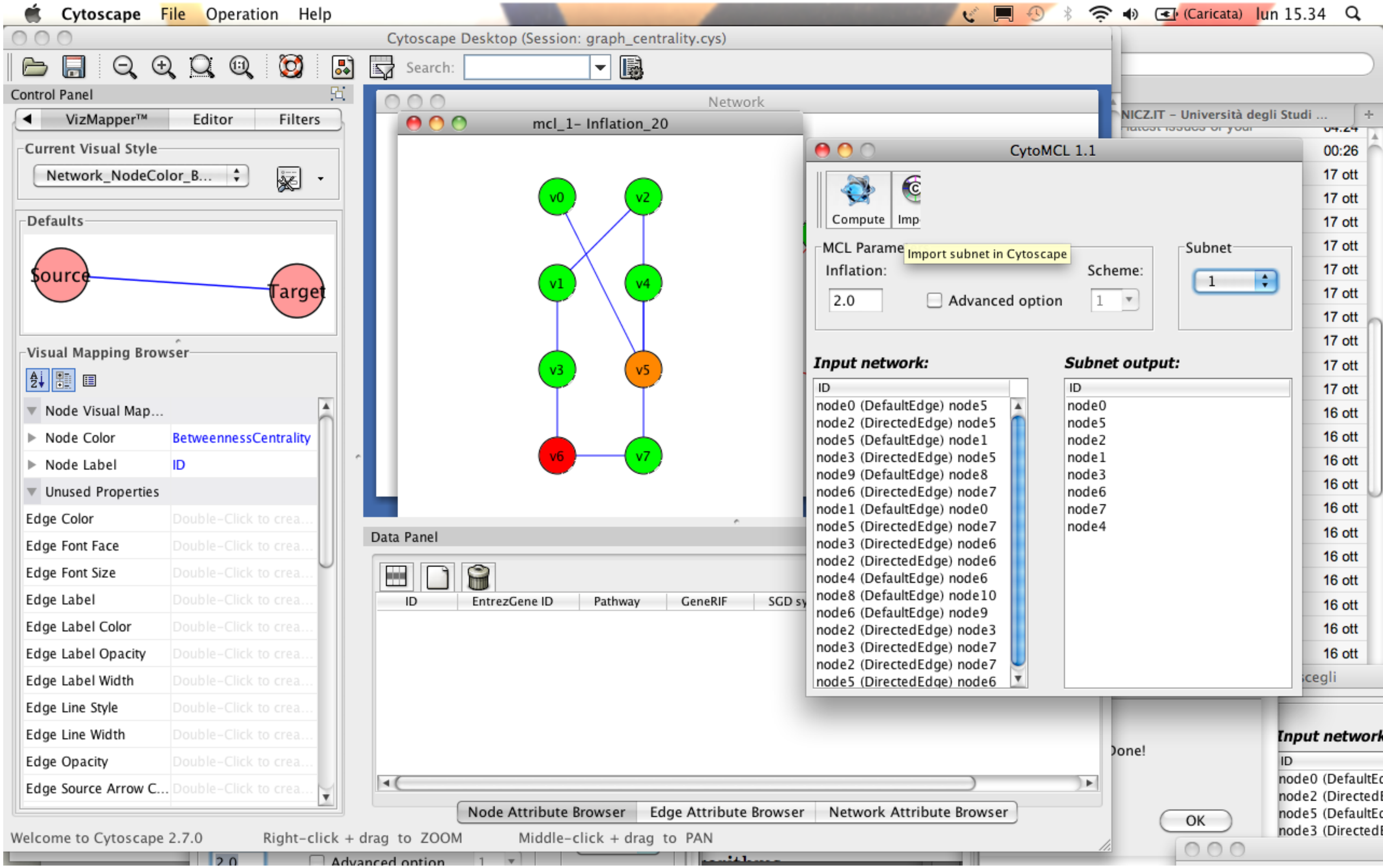
Input network:

ID
node0 (DefaultEdge) node5
node2 (DirectedEdge) node5
node5 (DefaultEdge) node1
node3 (DirectedEdge) node5
node9 (DefaultEdge) node8
node6 (DirectedEdge) node7
node1 (DefaultEdge) node0
node5 (DirectedEdge) node7
node3 (DirectedEdge) node6
node2 (DirectedEdge) node6
node4 (DefaultEdge) node6
node8 (DefaultEdge) node10
node6 (DefaultEdge) node9
node2 (DirectedEdge) node3
node3 (DirectedEdge) node7
node2 (DirectedEdge) node7
node5 (DirectedEdge) node6

Subnet output:

ID
node0
node5
node2
node1
node3
node6
node7
node4

Using CytoMCL results in Cytoscape



Input network:

```
ID
node0 (DefaultEdge) node5
node2 (DirectedEdge) node5
node5 (DefaultEdge) node1
node3 (DirectedEdge) node5
node9 (DefaultEdge) node8
node6 (DirectedEdge) node7
node1 (DefaultEdge) node0
node5 (DirectedEdge) node7
node3 (DirectedEdge) node6
node2 (DirectedEdge) node6
node4 (DefaultEdge) node6
node8 (DefaultEdge) node10
node6 (DefaultEdge) node9
node2 (DirectedEdge) node3
node3 (DirectedEdge) node7
node2 (DirectedEdge) node7
node5 (DirectedEdge) node6
```

Subnet output:

```
ID
node0
node5
node2
node1
node3
node6
node7
node4
```

UMG

Dubium sapientiae initium

CytoSeVis: Semantic Similarity-Based Visualization of PPI Networks

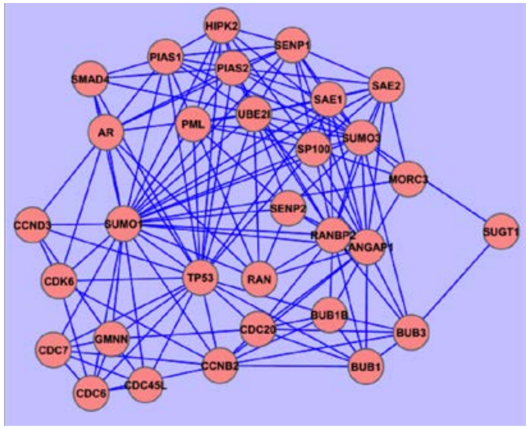
Mario Cannataro, Pietro Hiram Guzzi

University "Magna Graecia" of Catanzaro, Italy

cannataro@unicz.it

PINs Visualization

- PINs visualization tools should:
  - handle high-dimensional PINs
  - meet users' requests in terms of low response time, interactivity, ease of use
  - support some form of graph analysis

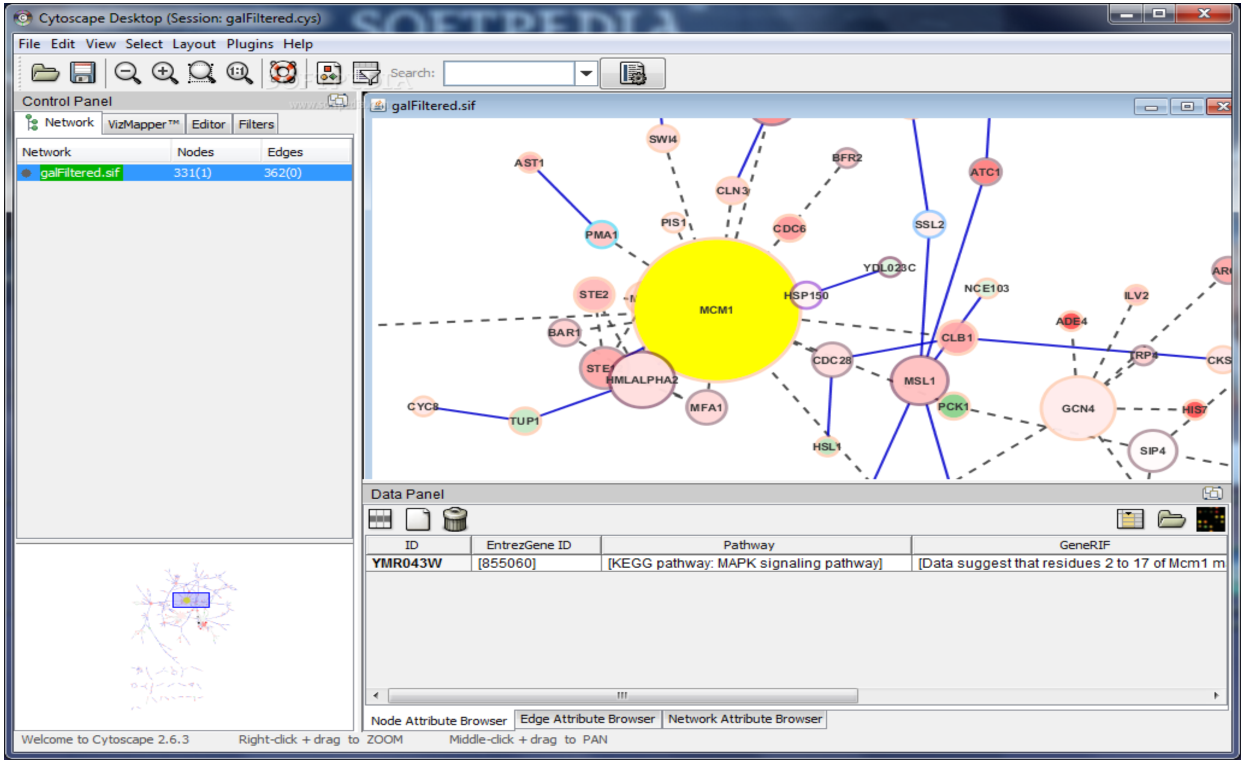


- PINs visualization tools comprise:
  - Collections of **layout algorithm** (e.g. circular, tree, hierarchical, Force-directed);
  - Different graphical **rendering algorithms** (2D, 3D)
  - **Graph-analysis** and **annotation** (clustering, statistics)
- Main tools include Cytoscape and NAViGaTOR

Agapito et al. Visualization of protein interaction networks: problems and solutions, BMC Bioinformatics 2013

Pavlopoulos et al. BioData Mining, 2008

Visualization in Cytoscape



Network	Nodes	Edges
galFiltered.sif	331(1)	362(0)

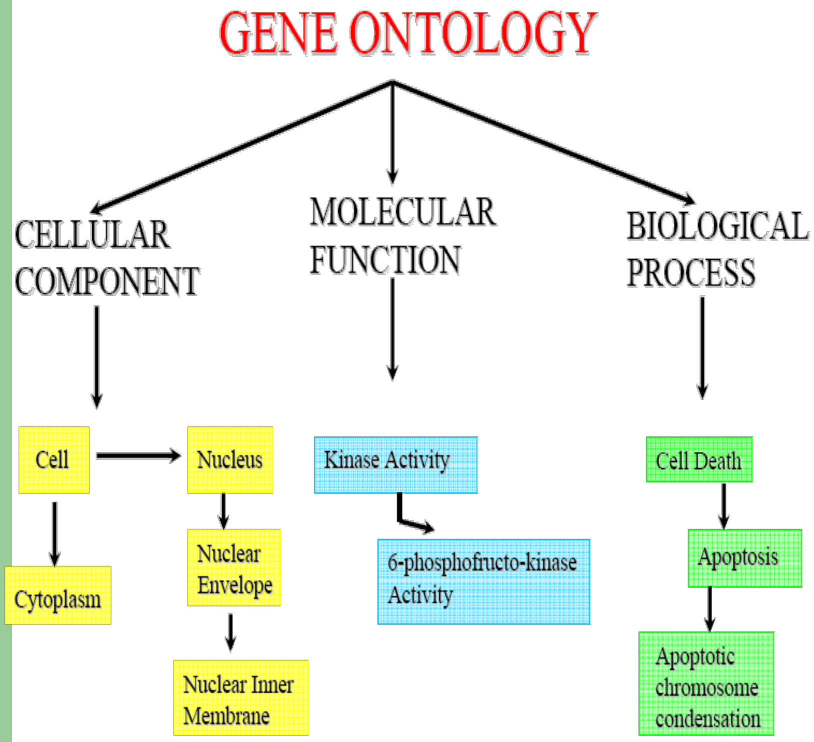
ID	EntrezGene ID	Pathway	GeneRIF
YMR043W	[855060]	[KEGG pathway: MAPK signaling pathway]	[Data suggest that residues 2 to 17 of Mcm1 m

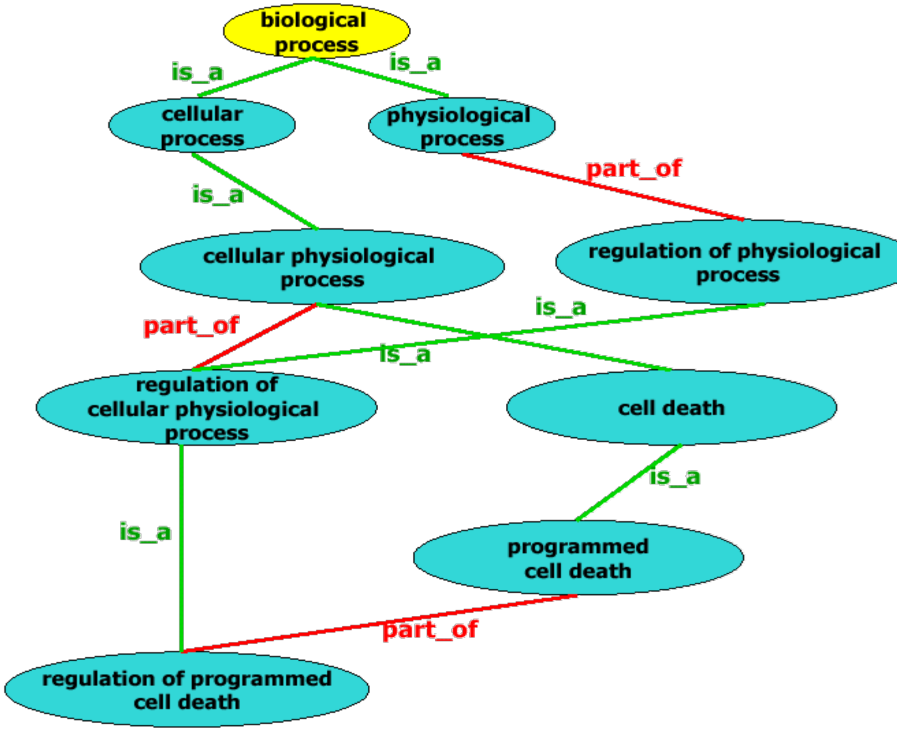
Welcome to Cytoscape 2.6.3 Right-click + drag to ZOOM Middle-click + drag to PAN

- **Layouts**: Spring-Embedded, Circular, Grid, Force-Directed and Organic;
- **Rendering**: 2D
- **Network analysis**: basic statistical analysis:
  - through the plug-in manager it is possible to add external modules
- **Cytoscape does not support full semantic similarity analysis**
  - GOlorize provides class-guided network visualization (used with BINGO)
  - BINGO finds GO categories overrepresented in a selected part of the network

BIOLOGICAL ONTOLOGIES

Biological Knowledge is encoded into many Biological Ontologies




biological process
is_a
is_a
cellular process
physiological process
part_of
is_a
cellular physiological process
regulation of physiological process
is_a
part_of
is_a
regulation of cellular physiological process
cell death
is_a
is_a
programmed cell death
part_of
regulation of programmed cell death

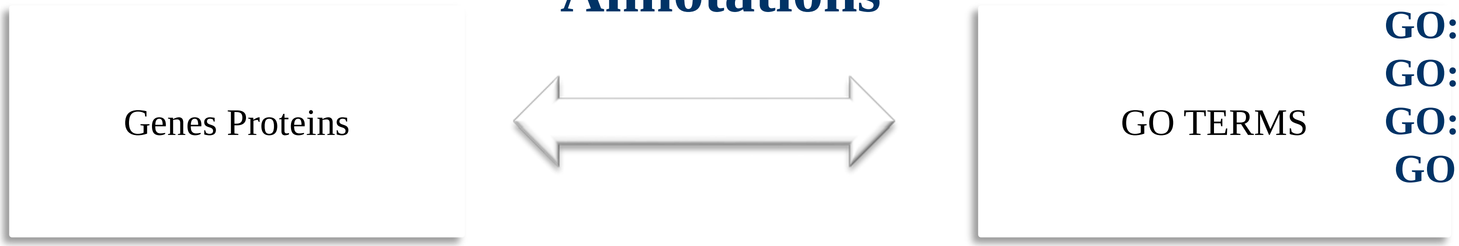
Semantic Similarities

- Biological entities and molecules are also associated with GO terms representing their functions, biological roles and localization.
- The association of GO terms and biological molecules is called **annotation process** and it may be performed manually under the supervision of an expert, or automatically.
- There exist currently 17 different annotation processes that are identified by an **evidence code.**
- The **set of annotations** is stored in different databases, such as *the Gene Ontology Annotation Database* (GOA)

Guzzi et al, Briefings in Bioinformatics, 2012

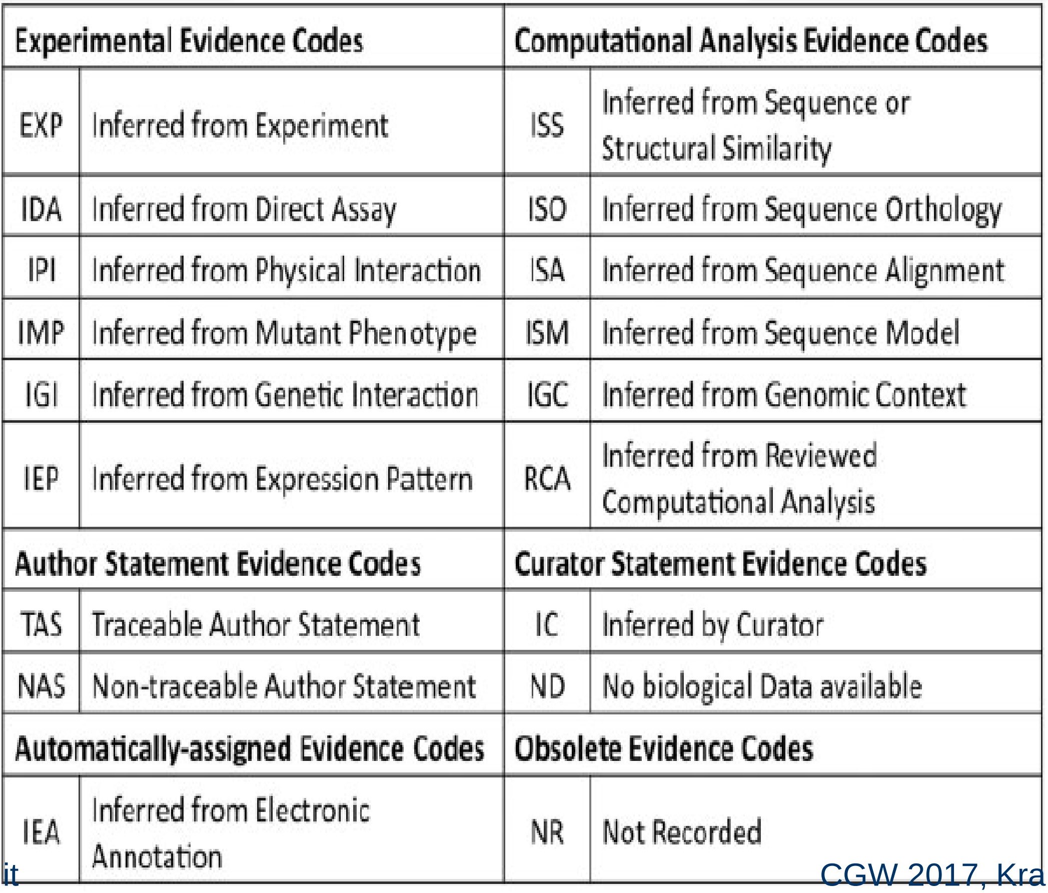
ANNOTATION PROCESS

Annotations


Genes Proteins

GO TERMS

GO:0008152,
GO:0004807,
GO:0016853,
GO:0003824



Experimental Evidence Codes		Computational Analysis Evidence Codes	
EXP	Inferred from Experiment	ISS	Inferred from Sequence or Structural Similarity
IDA	Inferred from Direct Assay	ISO	Inferred from Sequence Orthology
IPI	Inferred from Physical Interaction	ISA	Inferred from Sequence Alignment
IMP	Inferred from Mutant Phenotype	ISM	Inferred from Sequence Model
IGI	Inferred from Genetic Interaction	IGC	Inferred from Genomic Context
IEP	Inferred from Expression Pattern	RCA	Inferred from Reviewed Computational Analysis
Author Statement Evidence Codes		**Curator Statement Evidence Codes**	
TAS	Traceable Author Statement	IC	Inferred by Curator
NAS	Non-traceable Author Statement	ND	No biological Data available
Automatically-assigned Evidence Codes		**Obsolete Evidence Codes**	
IEA	Inferred from Electronic Annotation	NR	Not Recorded

Ontology-based analysis

- GO and GO annotations enable the use of a set of **analysis methodologies** that have been **defined for the ontologies.**

- There exists, in fact, a set of analysis methodologies of the terms of ontologies as well as of the annotated entities

- An important **tool for ontology-based analysis** is represented by the **pairwise semantic similarities measures (SSM).**
  - **SSM** are mathematical functions that quantify the similarities of two terms belonging to an ontology into a numerical value.
  - **SSM** are usually **defined for the comparison of two terms** but they can be easily extended to consider group of terms as input.

Semantic Similarity

- In computational biology, semantic similarity measures are generally based on Gene Ontology and are often used to compare both GO terms and gene products.

- There exist currently many available semantic similarity measures as listed in Pesquita et al. that can be categorized using different parameters on the basis of the steps they employ to determine the semantic value.

SEMANTIC SIMILARITIES

- **Pairwise: Two Terms**
- **Groupwise: Set of terms, i.e. proteins annotated with a set of terms**

Proteins are described with a set of annotating terms

TPI1 → GO:0008152,GO:0004807,GO:0016853, GO:0003824

PGAM2 → GO:0006096,GO:0008152,GO:0016853, GO:0003824,GO:0016868

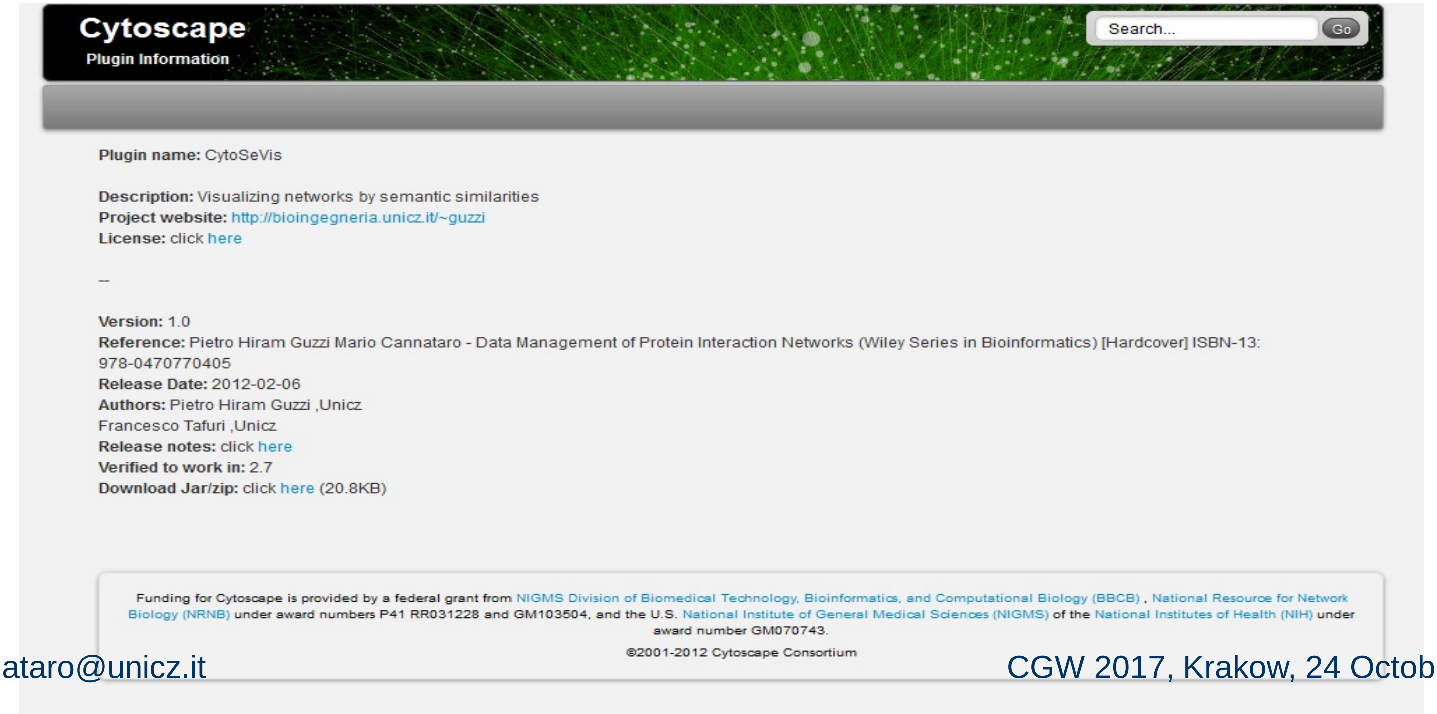
MEASURE	METHOD
Resnik	Pairwise
ResnikGraSM	Pairwise
Lin	Pairwise
LinGraSM	Pairwise
JiangConrath	Pairwise
JiangConrathGraSM	Pairwise
Relevance	Pairwise
Kappa	Groupwise
Cosine	Groupwise
SimGIC	Groupwise
CzekanowskiDice	Groupwise

Cyto-SeVis

- CytoSeVis is a Cytoscape plugin that is able to visualize protein interaction networks in a semantic similarity space
- CytoSevis is based on an intuitive interface and is able to load a network from Cytoscape
- It loads the semantic similarities provided as separate files and visualizes resulting coloured network into Cytoscape workspace

CytoSeVis Availability

- CytoSeVis for Windows, Linux and Mac OSX platforms is available under GPL license.
- Download at Cytoscape web site: http://cytoscape.org
- Semantic similarities for yeast available at:http://bioingegneria.unicz.it/~guzzi/ss



Plugin name: CytoSeVis

Description: Visualizing networks by semantic similarities
Project website: http://bioingegneria.unicz.it/~guzzi
License: click here

--

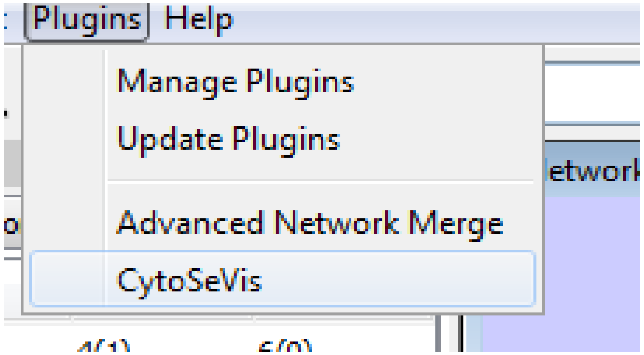
Version: 1.0
Reference: Pietro Hiram Guzzi Mario Cannataro - Data Management of Protein Interaction Networks (Wiley Series in Bioinformatics) [Hardcover] ISBN-13: 978-0470770405
Release Date: 2012-02-06
Authors: Pietro Hiram Guzzi ,Unicz
Francesco Tafuri ,Unicz
Release notes: click here
Verified to work in: 2.7
Download Jar/zip: click here (20.8KB)

Funding for Cytoscape is provided by a federal grant from NIGMS Division of Biomedical Technology, Bioinformatics, and Computational Biology (BBCB) , National Resource for Network Biology (NRNB) under award numbers P41 RR031228 and GM103504, and the U.S. National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under award number GM070743.

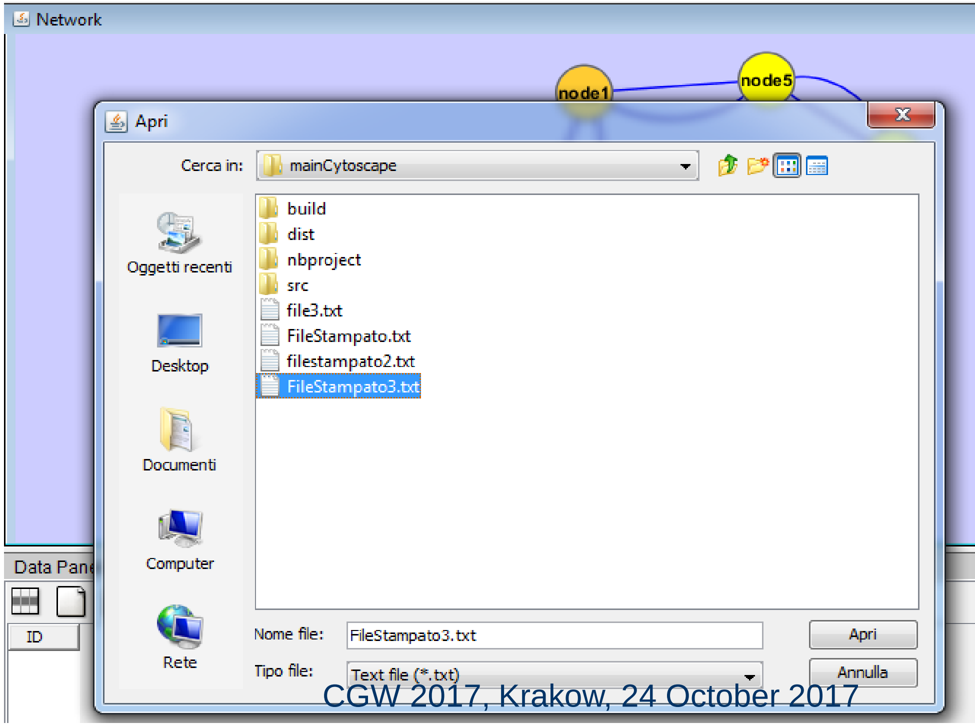
CytoSevis Operation

Once that a network has been loaded.

1. User has to select the CytoSeVis plugin and an initial node.



2. Then user has to select and load the similarity file



Cerca in: mainCytoscape

Nome file: FileStampato3.txt

Tipo file: Text file (*.txt)

CytoSeVis SS input file

SS input file is a NxN precomputed matrix containing SS among each couple of proteins

SS are calculated with csbl.go R package: http://csbi.ltdk.helsinki.fi/csbl.go/ (Ovaska et al, BioData Mining, 2008)

Available SS:

Cosine

Czekanowski Dice

JiangConrath

Kappa

Lin

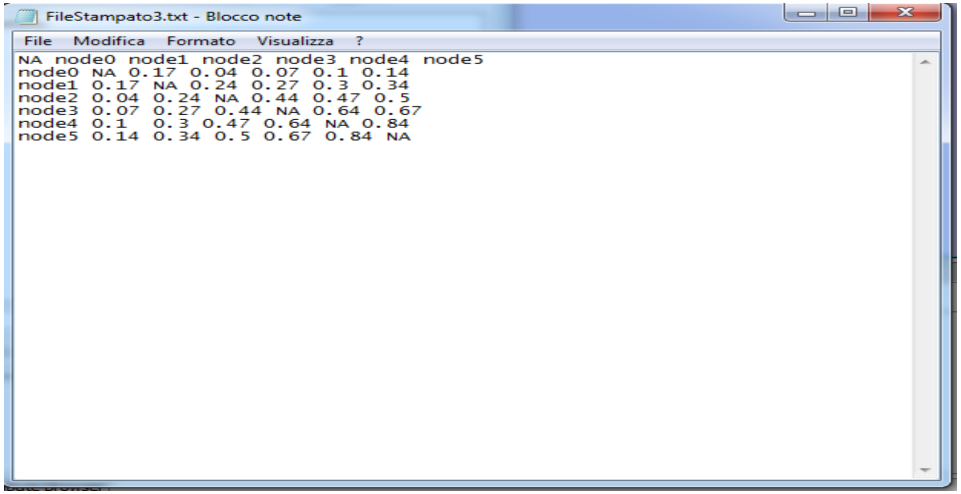
LinGrasm

Relevance

Resnik

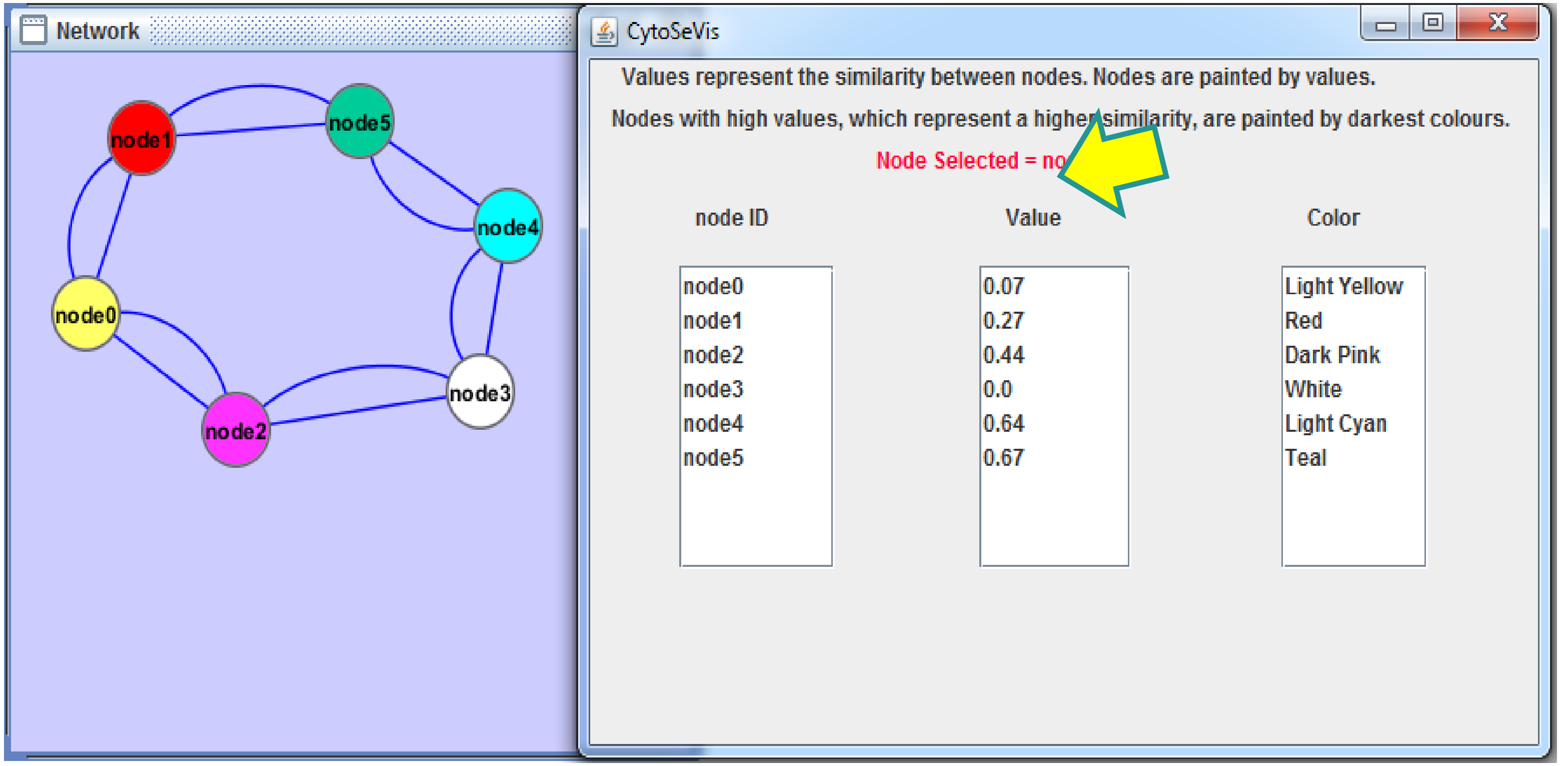
Weighted Jaccard

Resnik Grasm


FileStampato3.txt - Blocco note
File Modifica Formato Visualizza ?

```
NA node0 node1 node2 node3 node4 node5
node0 NA 0.17 0.04 0.07 0.1 0.14
node1 0.17 NA 0.24 0.27 0.3 0.34
node2 0.04 0.24 NA 0.44 0.47 0.5
node3 0.07 0.27 0.44 NA 0.64 0.67
node4 0.1  0.3 0.47 0.64 NA 0.84
node5 0.14 0.34 0.5 0.67 0.84 NA
```

CytoSeVis Use



Values represent the similarity between nodes. Nodes are painted by values.

Nodes with high values, which represent a higher similarity, are painted by darkest colours.

Node Selected = no

node ID	Value	Color
node0	0.07	Light Yellow
node1	0.27	Red
node2	0.44	Dark Pink
node3	0.0	White
node4	0.64	Light Cyan
node5	0.67	Teal

- Semantic similarities are translated into different colors.
  - The white node is the reference seed node.
  - All the other similarities are calculated with respect to this node.
- In the example, a darker color represents a lower similarity.

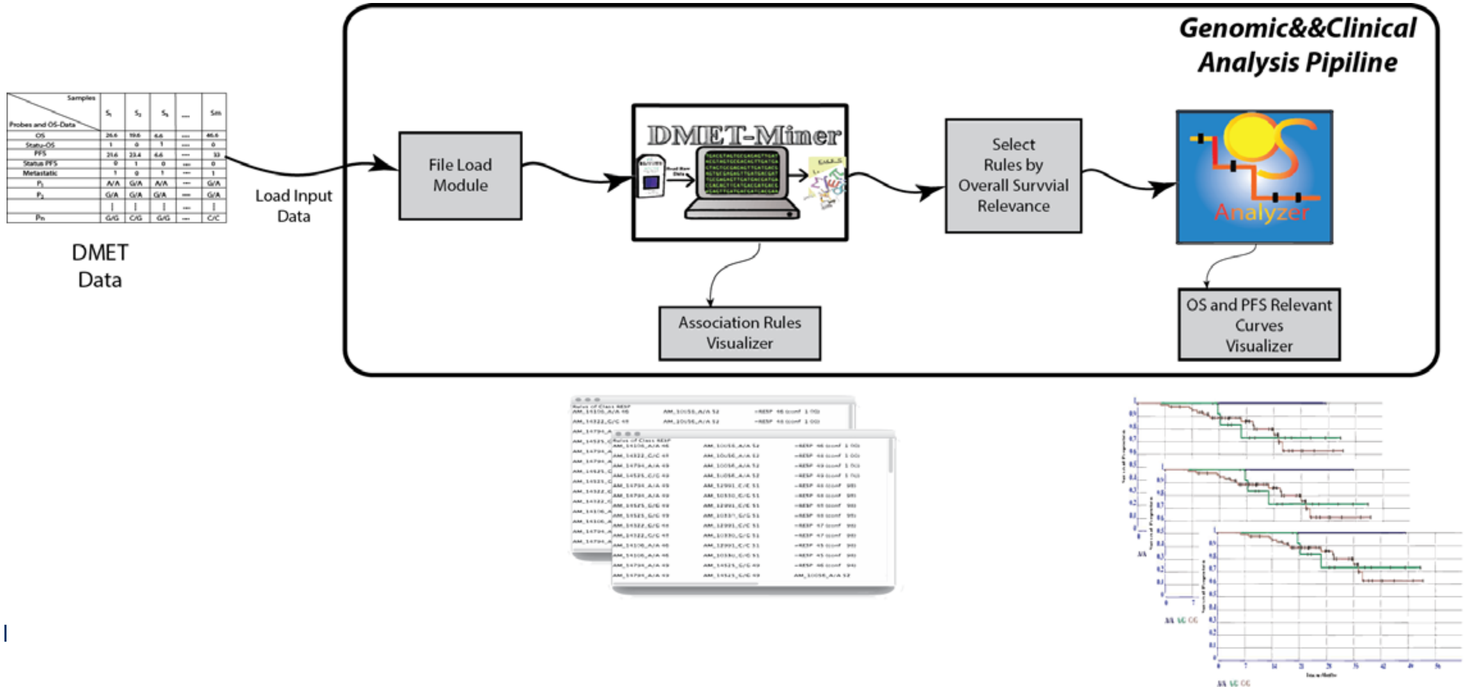
- Semantic similarity adds a new dimension for PINs visualization and analysis
- CytoSeVis is a Cytoscape plugin for semantic similarity-based visualization of PINs
  - http://cytoscape.org
- Semantic similarities are available at:http://bioingegneria.unicz.it/~guzzi/ss
- See also:
  - Guzzi et al Briefings in Bioinformatics 2012,
  - Guzzi et al IEEE BIBMW 2011, ICIAP 2011

OUTLINE

- Experiences at University of Catanzaro in high performance management, preprocessing and analysis of omics data
  - PART I: Genomics data
  - PART II: Proteomics data
  - PART III: Interactomics data
- **Conclusions**

Conclusions: toward an integrated pipeline

- **DMET-Analyzer** supports the automatic statistical analysis in DMET-based pharmacogenomics data and it is able to find statistically relevant subsets of SNPs that separate two input classes
- **DMET-Miner** extracts Association Rules from DMET-based pharmacogenomics data
- **OSAnalyzer** analyzes genomics data annotated with clinical data and computes the OS and PFS curves related to the presence/absence of SNPs in the population under investigation



Acknowledgements

- University of Catanzaro Clinical and Biology Units
  - U.O. Oncologia Policlinico Universitario Mater Domini
    - Prof. Pierosandro Tagliaferri, Prof. Pierfrancesco Tassone, Dr.ssa Maria Teresa Di Martino
  - Laboratorio di Proteomica e Spettrometria di Massa dell’Università di Catanzaro
    - Prof. Giovanni Cuda, Dr. Marco Gaspari
  - U.O. Biochimica Clinica Policlinico Universitario Mater Domini
    - Prof. Francesco Saverio Costanzo, Dr.ssa Barbara Quaresima, Dr.ssa Concetta Faniello
- ICAR-CNR and University of Calabria
  - Prof. Domenico Talia
- Georgia Institute of Technology and University of Padova
  - Prof. Concettina Guerra
- University of Salerno
  - Prof. Alessandro Weisz
- Fondazione Bruno Kessler (Trento)
  - Dr. Cesare Furlanello
- University of Calabria
  - Prof. Sergio Greco

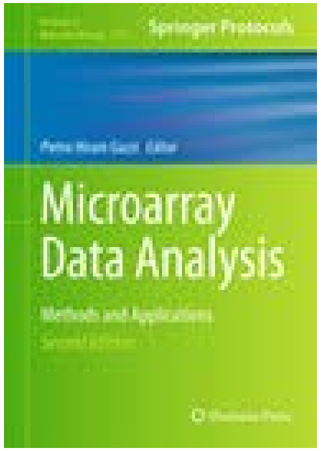
UMG

Dubium sapientiae initium

Bioinformatics Group - www.bioingegneria.it/~cannataro

- **Prof. Mario Cannataro**
- **Prof. Pierangelo Veltri**
- **Prof. Pietro H. Guzzi**

- Dr. Agapito Giuseppe, PhD, PostDoc;
- Dr. Calabrese Barbara, PhD, PostDoc;
- Dr. Giuseppe Tradigo, PhD, PostDoc;
- Dr. Vizza Patrizia, PhD, PostDoc;
- Ing. Cristiano Francesca, PhD Student;
- Ing. Milano Marianna, PhD Student;
- Ing. Mirarchi Domenico, PhD Student;
- Dr. Chiara Zucco, PhD Student;
- Ing. Pietro Cinaglia, PhD Student.



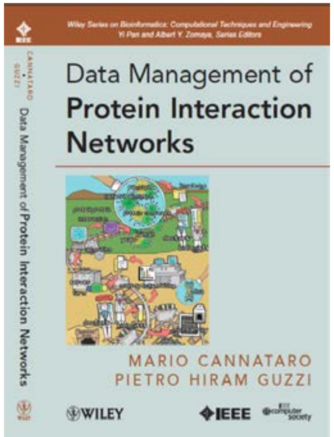
Springer Protocols

Pietro Hiram Guzzi Editor

Microarray Data Analysis

Methods and Applications

Second Edition



Wiley Series on Bioinformatics: Computational Techniques and Engineering

Yi Pan and Albert Y. Zomaya, Series Editors

Data Management of Protein Interaction Networks

MARIO CANNATARO

PIETRO HIRAM GUZZI

WILEY

IEEE

Briefings in Bioinformatics

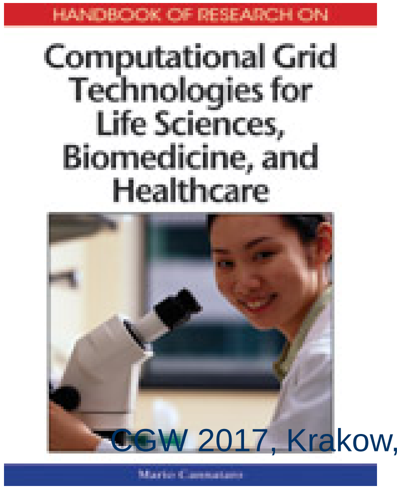
MARCH 2008 VOLUME 9 NUMBER 2

www.bib.oxfordjournals.org

SPECIAL ISSUE: COMPUTATIONAL PROTEOMICS

GUEST EDITOR: MARIO CANNATARO

OXFORD JOURNALS



HANDBOOK OF RESEARCH ON

Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare

Mario Cannataro