Statistics Applied to Bioinformatics

Multivariate analysis
Introduction

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Multivariate data

- Each row represents one object (also called unit)
- Each column represents one variable

	variable 1	variable 2	...	variable p
object 1	x_{11}	x_{21}	...	x_{p1}
object 2	x_{12}	x_{22}	...	x_{p2}
object 3	x_{13}	x_{23}	...	x_{p3}
object 4	x_{14}	x_{24}	...	x_{p4}
object 5	x_{15}	x_{25}	...	x_{p5}
object 6	x_{16}	x_{26}	...	x_{p6}
object 7	x_{17}	x_{27}	...	x_{p7}
object 8	x_{18}	x_{28}	...	x_{p8}
...	...	...	...	...
object n	x_{1n}	x_{2n}	...	x_{pn}

Multivariate data with an outcome variable

- The outcome variable (also called criterion variable) can be quantitative or nominal

	Predictor variables				**Criterion variable**
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
object 1	x_{11}	x_{21}	...	x_{p1}	y_1
object 2	x_{12}	x_{22}	...	x_{p2}	y_2
object 3	x_{13}	x_{23}	...	x_{p3}	y_3
object 4	x_{14}	x_{24}	...	x_{p4}	y_4
object 5	x_{15}	x_{25}	...	x_{p5}	y_5
object 6	x_{16}	x_{26}	...	x_{p6}	y_6
object 7	x_{17}	x_{27}	...	x_{p7}	y_7
object 8	x_{18}	x_{28}	...	x_{p8}	y_8
...	...	...	...	...	...
object n	x_{1n}	x_{2n}	...	x_{pn}	y_n

Typical questions in multivariate analysis

- No outcome variable
  - Can the objects be separated in distinct classes on the basis of the variables ?

    → **Cluster analysis**
  - Which variables, or combinations of variables (factors), are the most explanatory for the differences between objects ?

    → **Factor analysis**
- Quantitative outcome variable
  - Is the outcome variable correlated with the predictor variables ?

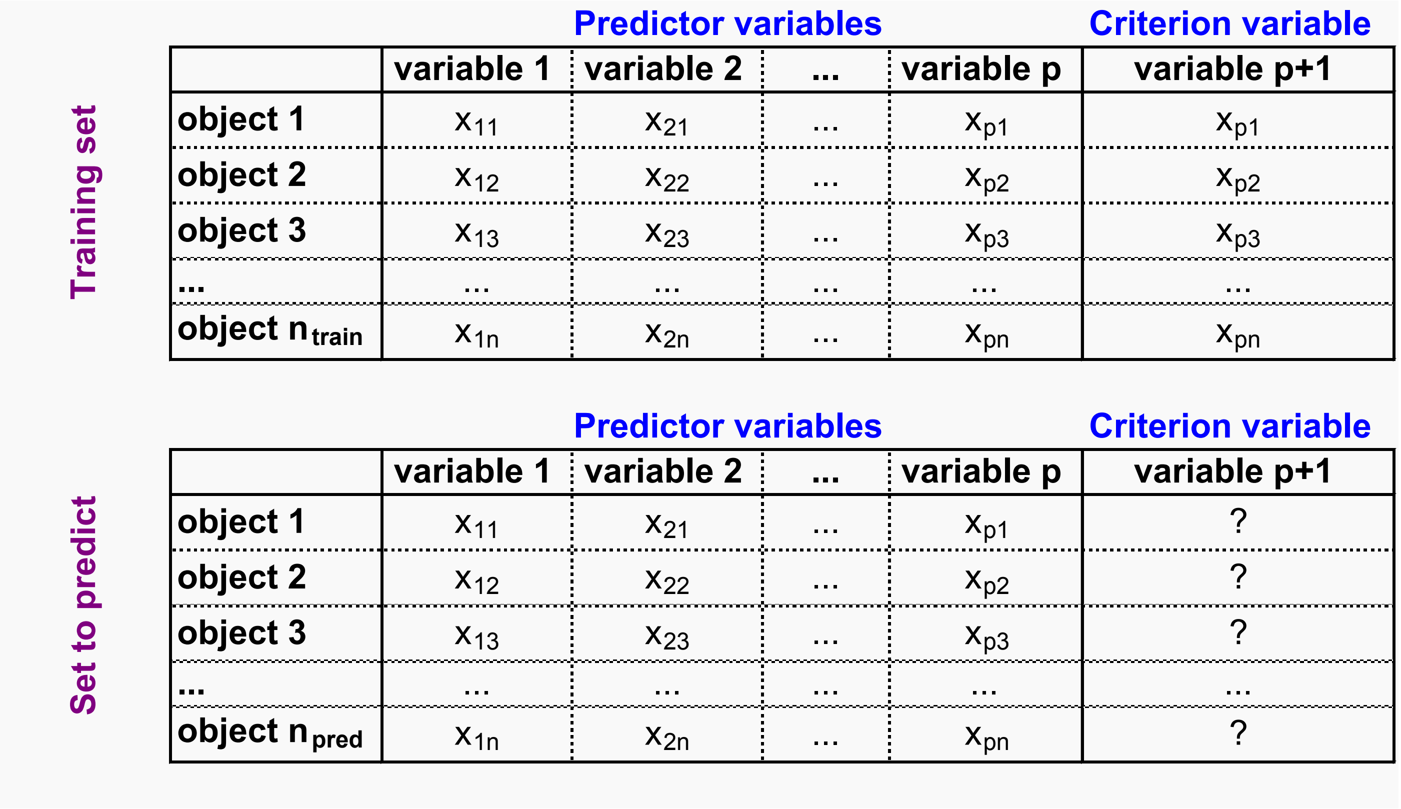
    → **Correlation analysis**
  - Can we predict the value of the outcome variable on the basis of the predictor variables ?

    → **Regression analysis**
- Nominal outcome variable
  - Can we predict the value of the outcome variable on the basis of the predictor variables ?

    → **Discriminant analysis**

Predictive approaches - Training set

- The training set is used to build a predictive function
- This function is used to predict the value of the outcome variable for new objects



Training set

	Predictor variables				Criterion variable
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
object 1	x_{11}	x_{21}	...	x_{p1}	x_{p1}
object 2	x_{12}	x_{22}	...	x_{p2}	x_{p2}
object 3	x_{13}	x_{23}	...	x_{p3}	x_{p3}
...	...	...	...	...	...
object n_{train}	x_{1n}	x_{2n}	...	x_{pn}	x_{pn}

Set to predict

	Predictor variables				Criterion variable
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
object 1	x_{11}	x_{21}	...	x_{p1}	?
object 2	x_{12}	x_{22}	...	x_{p2}	?
object 3	x_{13}	x_{23}	...	x_{p3}	?
...	...	...	...	...	...
object n_{pred}	x_{1n}	x_{2n}	...	x_{pn}	?

Evaluation of prediction with a testing set

Training set

	Predictor variables				Criterion variable
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
object 1	x_{11}	x_{21}	...	x_{p1}	x_{p1}
object 2	x_{12}	x_{22}	...	x_{p2}	x_{p2}
object 3	x_{13}	x_{23}	...	x_{p3}	x_{p3}
...	...	...	...	...	...
object n_{train}	x_{1n}	x_{2n}	...	x_{pn}	x_{pn}

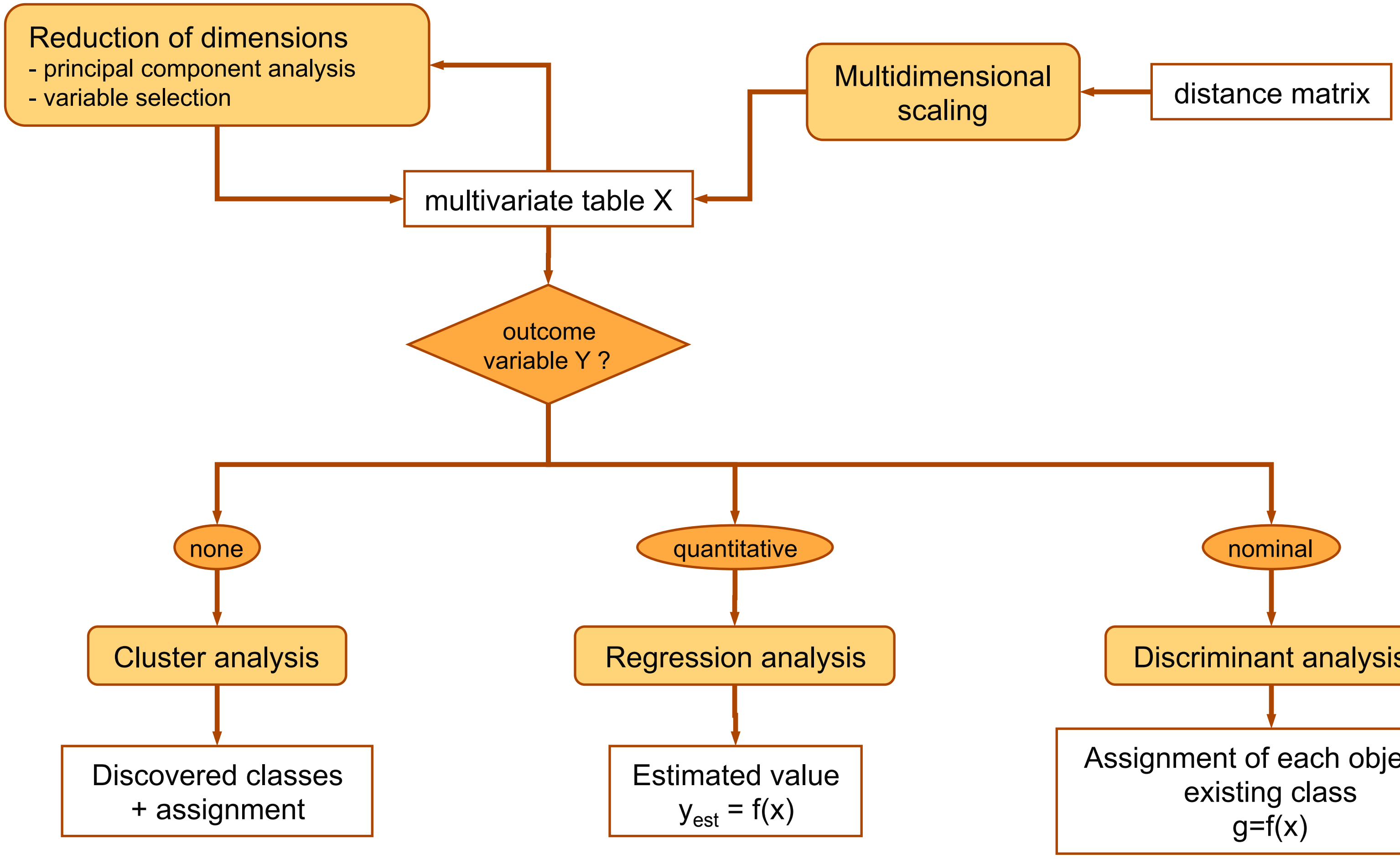
Testing set

	Predictor variables				Criterion variable	
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1 (known value)**	**variable p+1 (predicted)**
object 1	x_{11}	x_{21}	...	x_{p1}	x_{p1}	x'_{p1}
object 2	x_{12}	x_{22}	...	x_{p2}	x_{p2}	x'_{p2}
object 3	x_{13}	x_{23}	...	x_{p3}	x_{p3}	x'_{p3}
...	...	...	...	...	...	...
object n_{test}	x_{1n}	x_{2n}	...	x_{pn}	x_{pn}	x'_{p5}

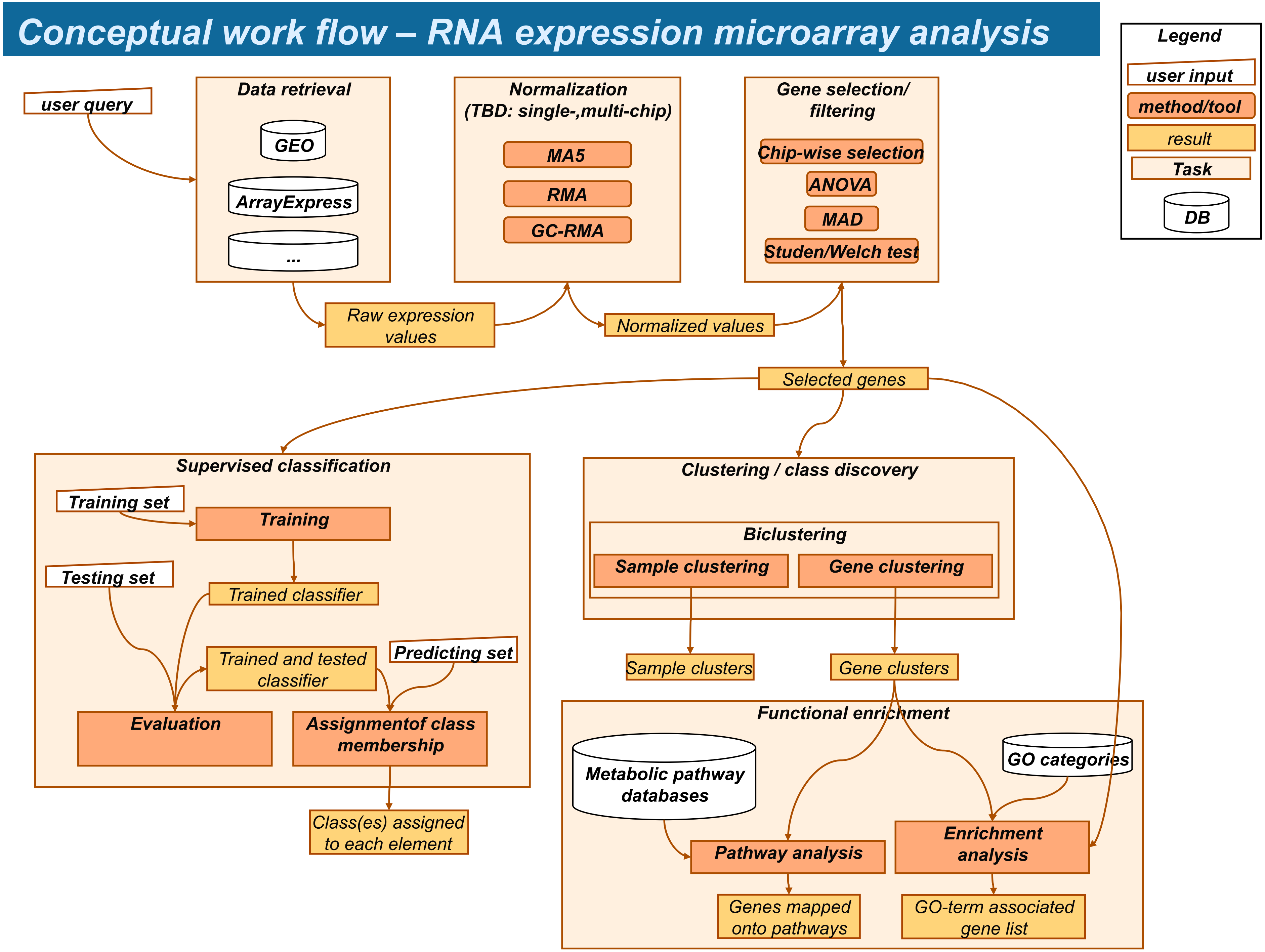
Set to predict

	Predictor variables				Criterion variable
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
object 1	x_{11}	x_{21}	...	x_{p1}	?
object 2	x_{12}	x_{22}	...	x_{p2}	?
object 3	x_{13}	x_{23}	...	x_{p3}	?
...	...	...	...	...	...
object n_{pred}	x_{1n}	x_{2n}	...	x_{pn}	?

Flowchart of the approaches in multivariate analysis



Conceptual work flow – RNA expression microarray analysis



Legend

user input

method/tool

result

Task

DB

user query

Data retrieval

GEO

ArrayExpress

...

Raw expression values

Normalization (TBD: single-,multi-chip)

MA5

RMA

GC-RMA

Normalized values

Gene selection/ filtering

Chip-wise selection

ANOVA

MAD

Studen/Welch test

Selected genes

Supervised classification

Training set

Training

Testing set

Trained classifier

Trained and tested classifier

Predicting set

Evaluation

Assignmentof class membership

Class(es) assigned to each element

Clustering / class discovery

Biclustering

Sample clustering

Gene clustering

Sample clusters

Gene clusters

Functional enrichment

Metabolic pathway databases

GO categories

Pathway analysis

Enrichment analysis

Genes mapped onto pathways

GO-term associated gene list