ROMA TRE
UNIVERSITÀ DEGLI STUDI

BIOMEDICAL DATA MANAGEMENT AND ANALYSIS

Emanuel Weitschek

emanuel@dia.uniroma3.it

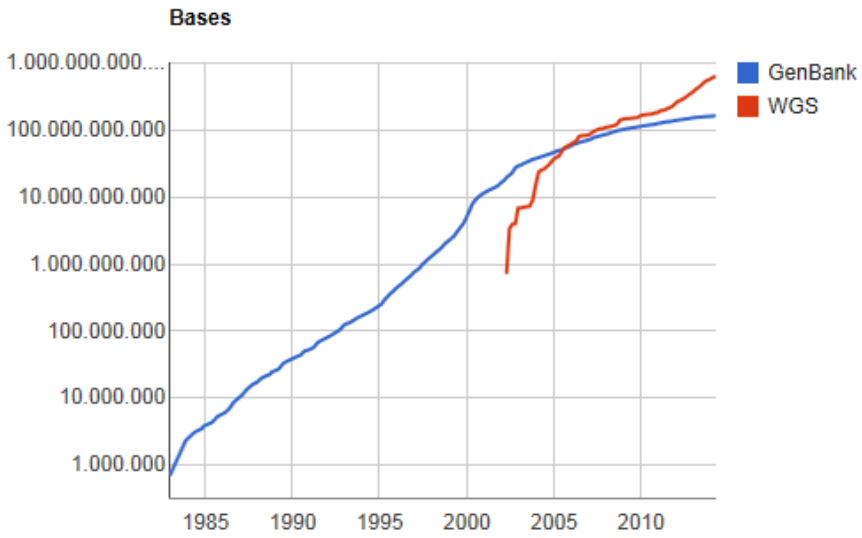
National Research Council of Italy

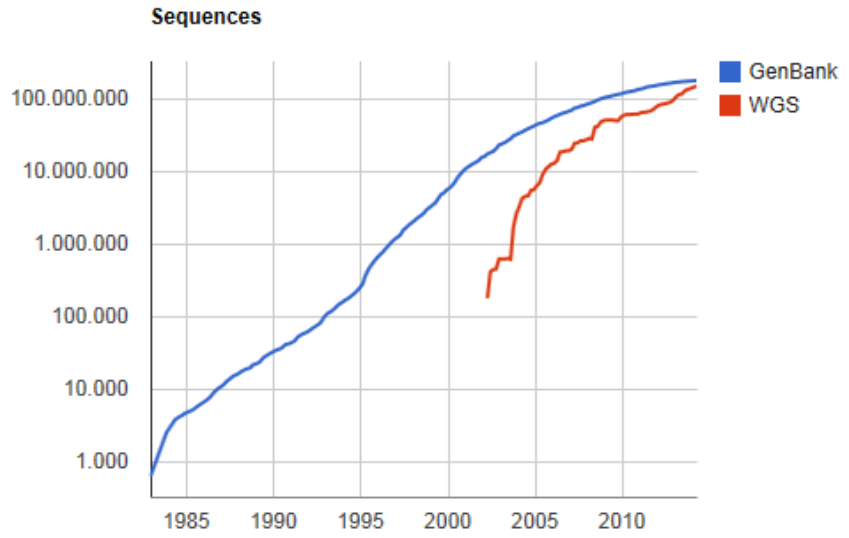
Outline

- Growth of biological data
- DNA Sequencing
- Growth of clinical data
- Analysis of biomedical data
- Bioinformatics
- GenData 2020
- The Cancer Genome Atlas and GenData
- Data Mining: classification and clustering
- Data mining systems: Weka and DMB
- Application to biological data sets
  - Sequences: DNA Barcode, Polyomaviruses, Bacteria and CNEs (with alignment free methods)
  - Gene Expression Profiles
  - Clinical patients
  - EEG signal processing
- Projects

Growth of biological data

- Advances in molecular biology lead to an exponential growth of biological data thanks to the support of computer science
  - originated by the DNA sequencing method invented by Sanger in early eighties
  - late nineties significant advances in sequence generation, e.g. Human Genome Project
  - actually the genomic sequences are doubling every 18 months
  - GenBank: collection of all publicly available nucleotide sequences (160 M seq)

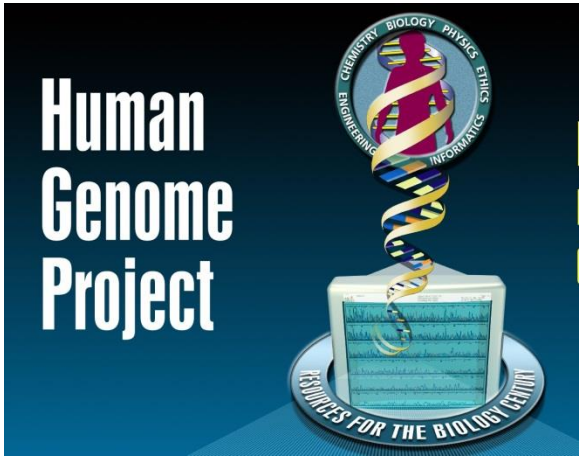


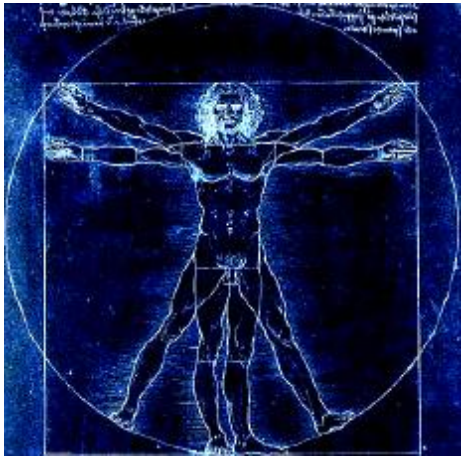


the era of "Big Data"

Growth of biological data

- Advances in molecular biology lead to an exponential growth of biological data thanks to the support of computer science
  - Today next generation high throughput data from modern parallel sequencing machines, are collected and huge amounts of biological data are currently available on public and private sources
  - 1000 Human Genomes project (3000 Mbp)
  - The future: 1000$ genome

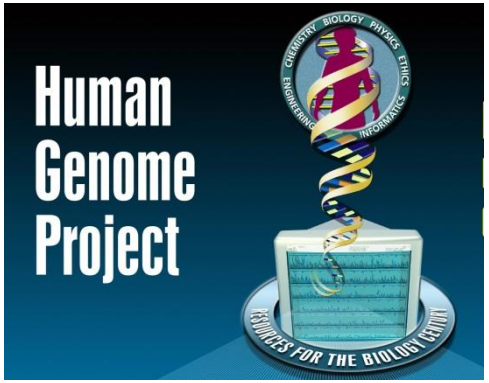


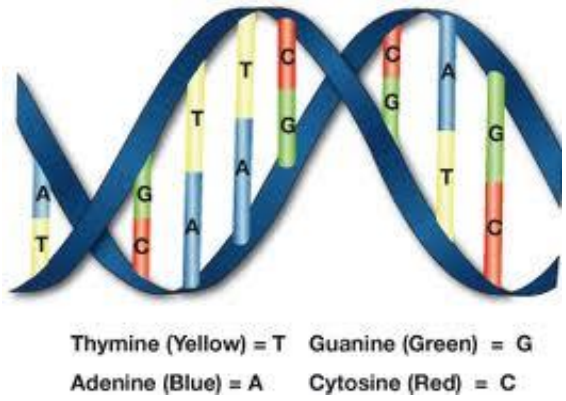


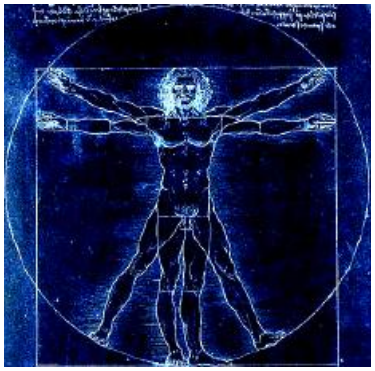
- Very large data sets, that are generated by several different biological experiments, need to be automatically processed and analyzed with computer science methods

DNA Sequencing

- DNA (deoxyribonucleic acid) is the hereditary material in almost all organisms
- DNA sequencing is the process of determining the order of nucleotides within a DNA molecule
- It includes any method or technology that is used to determine the order of the four bases—adenine (A), cytosine (C), guanine (G), and thymine (T)
- Originated by the DNA sequencing method invented by Sanger in early eighties
- In late nineties significant advances in sequence generation techniques, largely inspired by massive projects such as the Human Genome Project
- High costs and time, e.g., for the Human Genome Project 5 billions $ and 13 years
- Human Genome = 3 billion nucleotide long “book” written in A,C,G,T alphabet
- Different people = different genome (1 mutation each 1000 nt)
- Some species genome = 100x longer than human genome

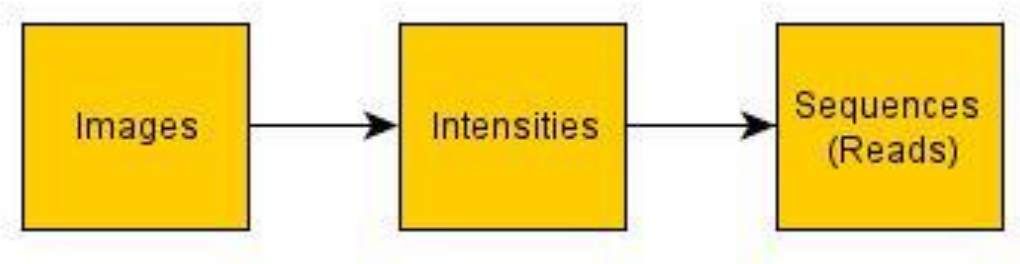






Next Generation Sequencing (NGS)

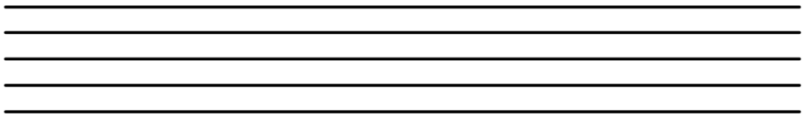
- Today: next generation high throughput data from modern parallel sequencing machines
  - Roche 454, Illumina, Applied Biosystems SOLiD, Helicos Heliscope, Complete Genomics, Pacific Biosciences SMRT, ION Torrent
  - Next generation sequencing (NGS) machines output a large amount of short DNA sequences, called reads (in fastq format)
  - Cannot read entire genome one nucleotides at a time from beginning to end
  - Can Shred genome and generate shorts reads
  - Low cost per base (5000$ for a whole human genome)
  - High speed (24h to sequence a whole human genome)
  - Large number of reads
  - Problems: data storage and analysis, high costs for IT infrastructure



ATGAAAAAGACAGCTATCGCGATTGCAGTGGCACTGGCTGGTTTCGCTACCGTGGCC
CAGGCGGCCTCTGAGGGAAACAGTGACTGCTACTTTGGGAATGGGTCAGCCTACCG
TGGCACGCACAGCCTCACCGAGTCGGGTGCCTCCTGCCTCCCGTGGAATTCCATGAT
CCTGATAGGCAAGGTTTACACAGCACAGAACCCCAGTGCCCAGGCACTGGGCCTGG
GCAAACATAATTACTGCCGGAATCCTGATGGGGATGCCAAGCCCTGGTGCCACGTG
CTGAAGAACCGCAGGCTGACGTGGGAGTACTGTGATGTGCCCTCCTGCTCCACCTGC
GGCCTGAGACAGTACAGCCAGCCTCAGTTTCGCATCAAAGGAGGGCTCTTCGCCGA
CATCGCCTCCCACCCCTGGCAGGCTGCCATCTTTGCCAAGCACAGGAGGTCGCCCGG
AGAGCGGTTCCTGTGCGGGGGCATACTCATCAGCTCCTGCTGGATTCTCTCTGCCGC
CCACTGCTTCCAGGAGAGGTTTCCGCCCCACCACCTGACGGTGATCTTGGGCAGAAC
ATACCGGGTGGTCCCTGGCGAGGAGGAGCAGAAATTTGAAGTCGAAAAATACATTG
TCCATAAGGAATTCGATGATGACACTTACGACAATGACATTGCGCTGCTGCAGCTGA
AATCGGATTCGTCCCGCTGTGCCCAGGAGAGCAGCGTGGTCCGCACTGTGTGCCTTC
CCCCGGCGGACCTGCAGCTGCCGGACTGGACGGAGTGTGAGCTCTCCGGCTACGGC
AAGCATCAGGCCTTGTCTCCTTTCTATTCGGAGCGGCTGAAGGAGGCTCATGTCAGA
CTGTACCCATCCAGCCGCTGCACATCACAACATTTACTTAACAGAACAGTCACCGAC
AACATGCTGTGTGCTGGAGACACTCGGAGCGGCGGCCCCAGGCAAACTTGCACGA
CGCCTGCCAGGGCGATTCGGGAGGCCCCCTGGTGTGTCTGAACGATGGCCGCATGA
CTTTGGTGGGCATCATCAGCTGGGGCCTGGGCTGTGGACAGAAGGATGTCCCGGGT
GTGTACACAAAGGTTACCAACTACCTAGACTGGATTCGTGACAACATGCGACCG
(SEQ ID No:2)

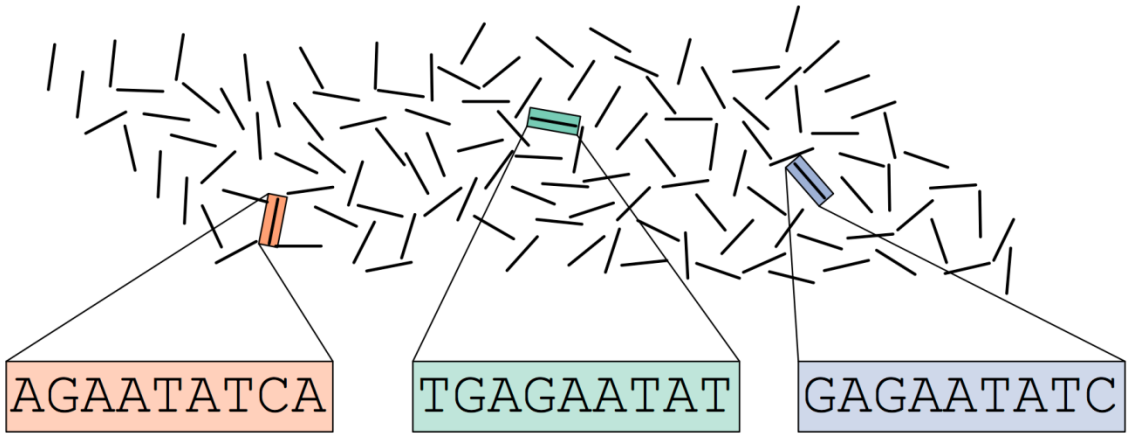
How do we assemble the genome?

Multiple copies of a genome



Shatter the genome into reads

Sequence the reads

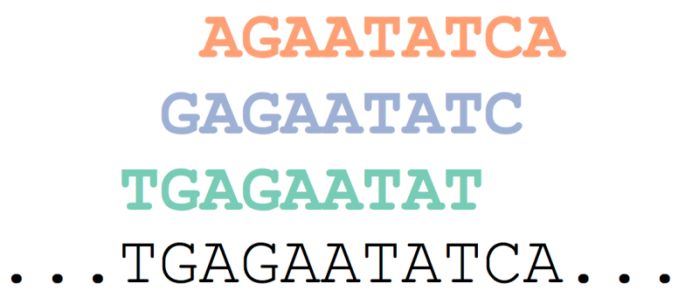


Assemble the genome with overlapping reads

AGAATATCA
GAGAATATC
TGAGAATAT
...TGAGAATATCA...

How do we assemble the genome?

- Take small tissue or blood sample containing millions of cells with identical DNA
- Use biochemical methods to break the DNA into fragments,
- Sequence these fragments to produce reads
- **Genome assembly**: putting a genome back together from its reads (it is just like reassembling a newspaper)
- Biologists can easily generate enough reads to analyze a large genome, but assembling these reads still presents a major computational challenge
- The difficulty is that researchers do not know where in the genome these reads came from, and so they must use overlapping reads to reconstruct the genome



NGS experiments

DNA-seq

- Sequence the DNA (whole genome) of an organism/sample
- De Novo assembly and Short read mapping
- Align the sequence to a reference genome (mutations analysis)

RNA-seq

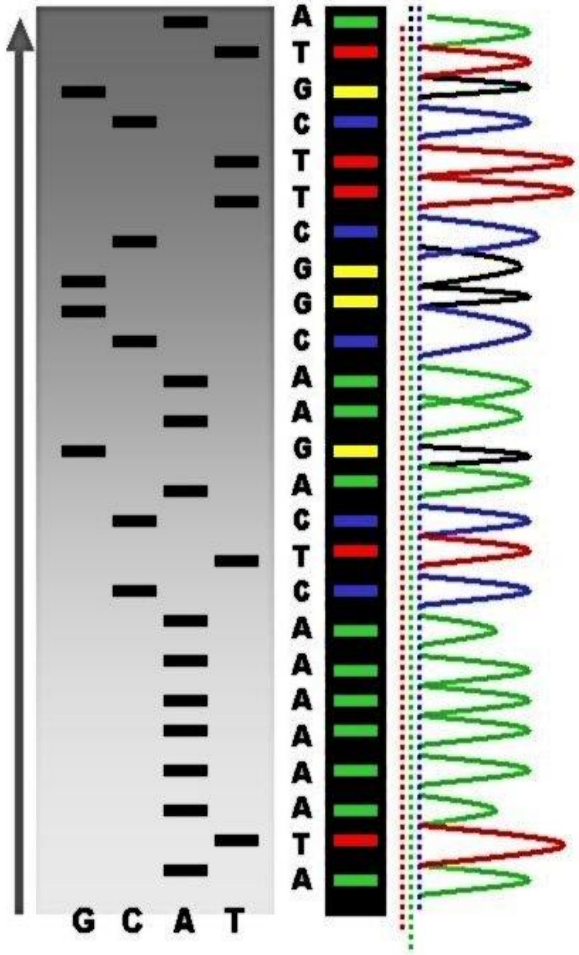
- Transcriptome sequencing to detect the gene expression profiles
- Count the reads that map on a particular gene region (quantitative measures)

Chip-seq

- Detection of genome-wide protein-DNA interactions
- Determines if a protein is bound to potential target regions

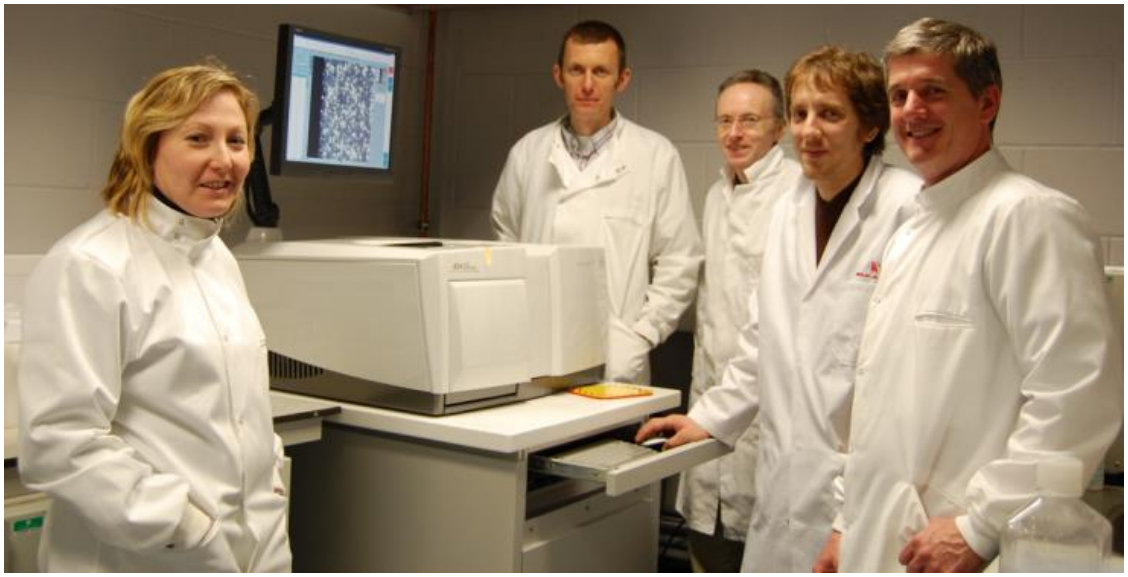
DNA methylation

- Determines epigenetic modifications which influence gene expression and cell phenotypes
- Epigenetics: study of changes in gene expression or cellular phenotype, caused by mechanisms other than changes in the underlying DNA sequence



Growth of clinical data

- The wide spread of electronic data collection in medical environments lead to an exponential growth of clinical data from heterogeneous patient samples
  - Electronic health records
  - European and national projects
  - Integration with genomics



the era of "Big Data"

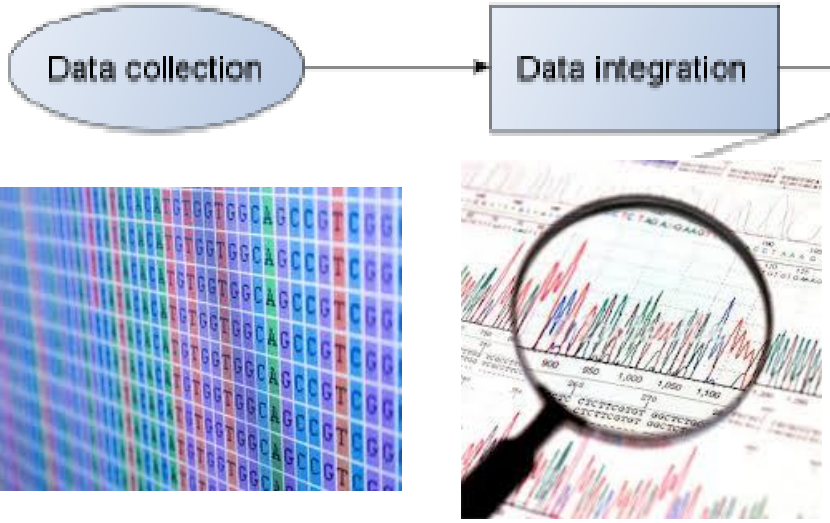
- Very large data sets, that are generated by several different clinical experiments, need to be automatically processed and analyzed with computer science methods

Growth of clinical data

- Collection of different electronic health patient records
- Data set obtained often from different health care facilities and hospitals
- Data collected by medical doctors or assistans: manual insertion of the clinical trials values or copy from non electronic medical records
- Normally no integrated IT system
- Therefore clinical data sets are often noisy, full of missing data and outliers
- To perform a data mining analysis the records have to be integrated in an unique data set
  - by joining on common attributes
  - by leaving all the non shared attributes as supplementary data
- Privacy issues:
  every patient must remain anonymous by mapping it with a private id, that has not to be visible to the data analyst

Analysis of biomedical data

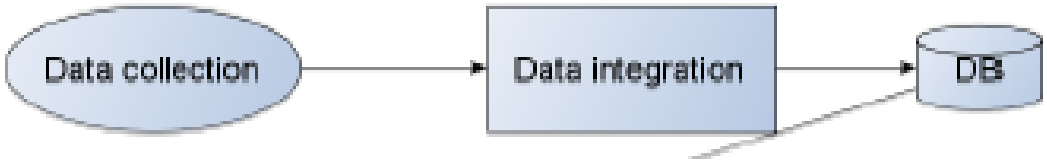
- Analyzing these enormous amount of data is becoming very important in order to shed light on biological and medical questions
- Challenges:
  - data collection
  - data integration
  - managing this huge amount of data
  - discovering the interactions
  - the integration of the biological know-how



Analysis of biomedical data

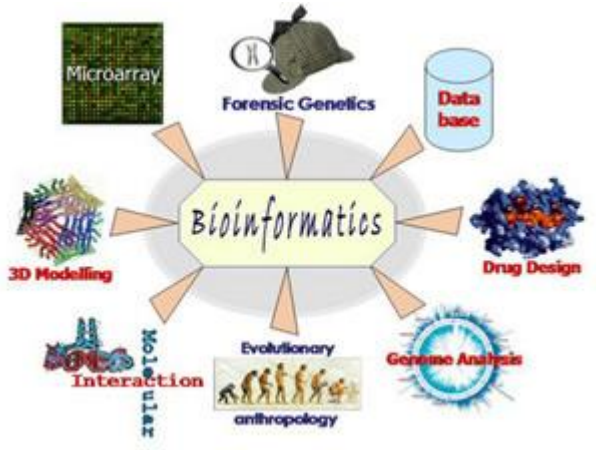
- Major issues:
  - incompleteness (missing values)
  - different adopted measure scales
  - integration of the disparate collection procedures
- Major problems in clinical databases
  - Complexity
  - Inaccuracy
  - Frequent missing values
  - Different adopted measure scales
  - Mixed (numeric – textual) variables, e.g. *BUN = {12, 16, high, 9, 11 …}*

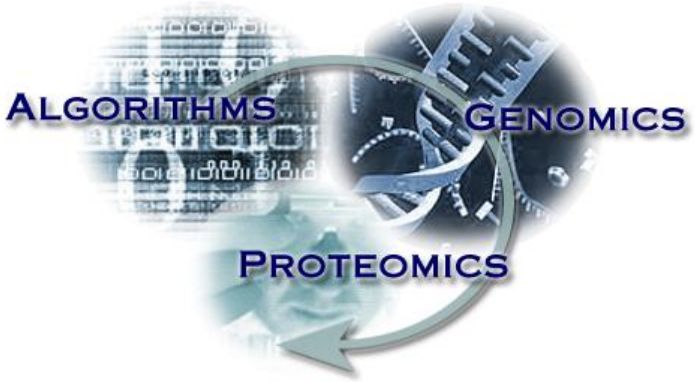
- **Final goal:** ***extract relevant information from huge amounts of biomedical data***

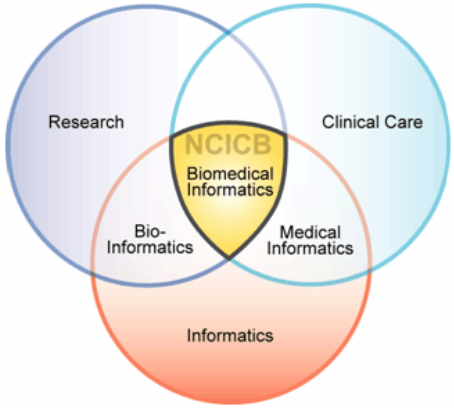


Bioinformatics

- New methods are demanded able to extract relevant information from biological data sets
- Effective and efficient computer science methods are needed to support the analysis of complex biological data sets
- Modern biology is frequently combined with computer science, leading to Bioinformatics
- **Bioinformatics** is a discipline where biology and computer science merge together in order to design and develop efficient methods for analyzing biological data, for supporting in vivo, in vitro and in silicio experiments and for automatically solving complex life science problems
- **Bioinformatician**: a computer scientist and biology domain expert, who is able to deal with the computer aided resolution of life science problems





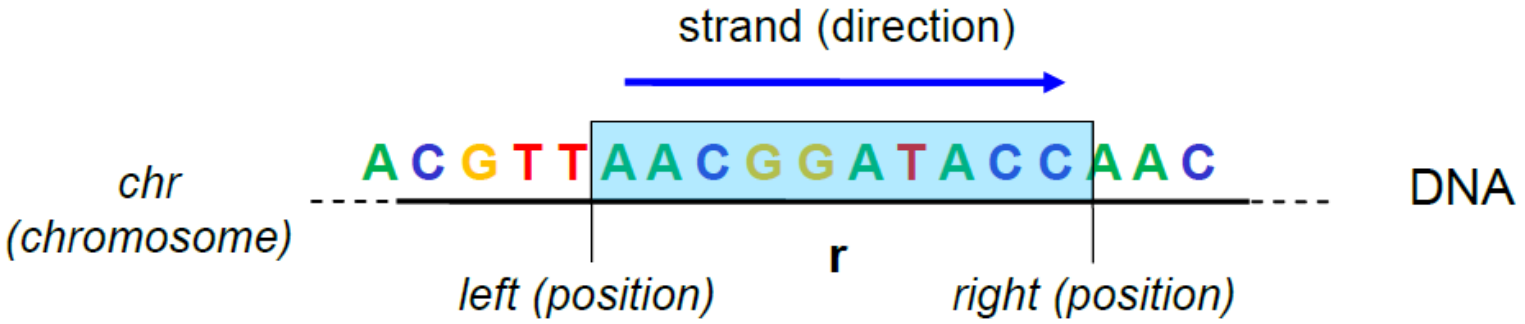


GenData 2020

- Data standard definition for genomics and biomedical data
- Build the abstractions, models, and protocols for supporting a network of genomic data
- Genome servers located in the major biologist laboratories in the world
- Distributed computing
- Definition of methods for querying, searching, and analyzing genomic data
- Problems: data management
  - huge amount of data
  - diversity of the platforms
  - diversity of the formats
- How to model and store genetic data so as to gain their integrated accessibility
- Data definition based on the standards proposed by the Functional Genomics Data Society
- Use of Distributed Annotation System (DAS) that defines a communication protocol used to exchange data on genomic or protein sequences

GenData 2020

- Final aim:
  Definition of an unique and global platform for effectively storing, searching and retrieving genomic data via distributed computing
- Focus NGS experiments:
  - DNA-seq, RNA-seq, ChIP-seq, DNA methylation
- GenData Format:
  - Within the same dataset, two kinds of data:
    - Region values aligned w.r.t. a given reference, with specific left-right ends within a chromosome



    - Metadata, with free-format attributes, storing all the knowledge about the dataset

- **Region values**: {*expID, region:(chr, left, right, strand), p-value*}

```
1    (3, 3245, 4535, +)        0.24 10^-8
1    (3, 5443, 6553, +)        0.44 10^-6
1    (3, 59873, 85443, *)      0.35 10^-8
1    (4, 653, 899, -)          0.43 10^-8
1    (15, 9874, 32345, +)      0.26 10^-7
2    (2, 586, 910, *)          0.51 10^-7
2    (5, 1274, 2421, -)        0.16 10^-8
2    (20, 35742, 39145, +)     0.57 10^-6
……
```

- **Metadata**: {*expID, attribute, value*}

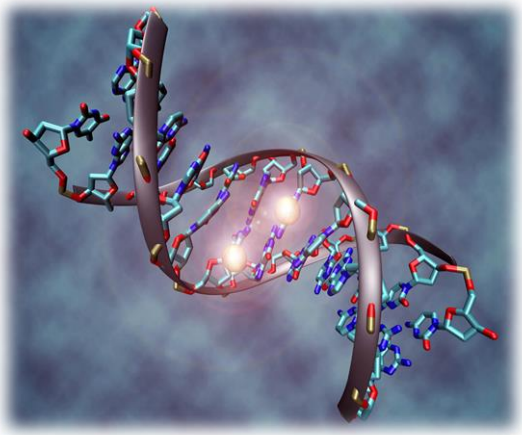
```
1    taxonomy     "Homo sapiens"
1    tissue       "Brain"
1    type         "ChIP-seq"
1    antibody     "cMyc"
2    taxonomy     "Homo sapiens"
2    tissue       "Breast"
2    type         "ChIP-seq"
……
```

GenData 2020

- Genometric Query Language (GMQL) is defined as a sequence of algebraic operations following the structure:
  < variable > = < operator > (< parameters >) < variable >
- Every variable is a collection of datasets describing experiments of the same level
- Offers high-level, declarative operations based on:
  - regions and their properties
  - general-purpose meta-data
- Inspired by Pig Latin and targeted towards cloud computing
- Will be further developed by adding genometric operations and statistical operations (e.g. clustering) in cooperation with biologists.
- The language allows for queries on the genome involving large datasets describing:
  - genomic signals
  - reference regions
  - distance rules

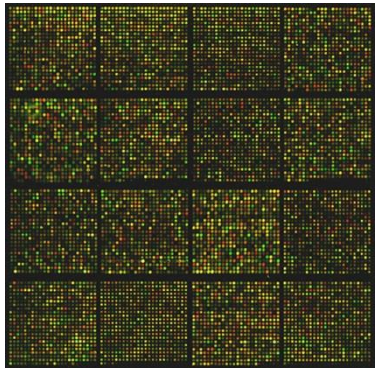
The Cancer Genome Atlas (TCGA)

- Comprehensive genomic characterization and analysis of more than 30 cancer type tissues to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing
- (2006) Coordinated joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) both of the National Institute of Health (NIH)
- **Aim:** improve the ability to diagnose, treat and prevent cancer
- A free-available platform to search, download, and analyze data sets containing clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes
- Open access



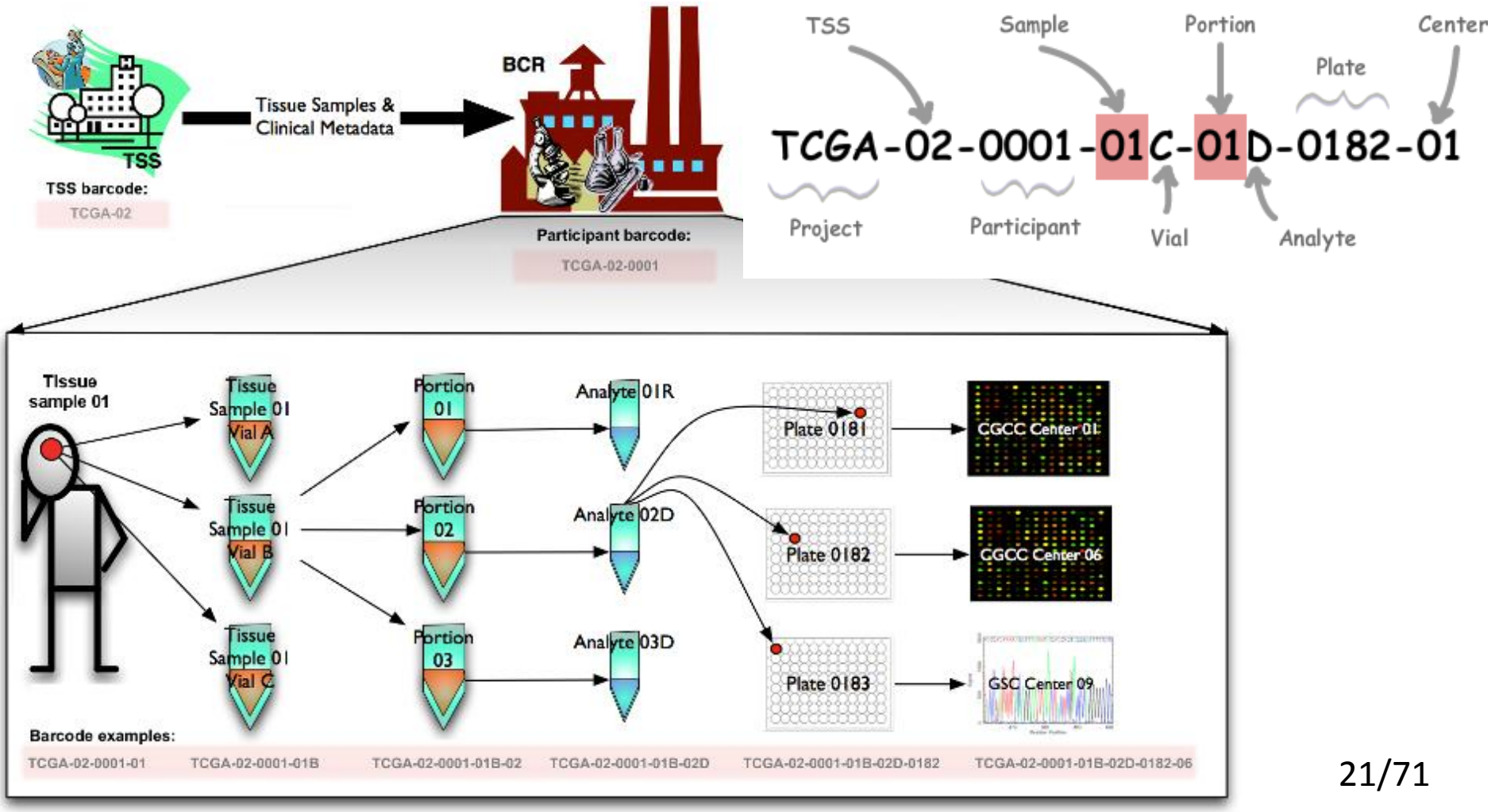
Data types

- Tissue pathology data
- Images
- Pathology reports
- Copy-number alterations for non-genetic platforms
- Epigenetic data
- Data summaries, such as genotype frequencies
- Clinical data (biotab and xml)
- DNA Sequencing (whole genome, whole exome, mutations)
- DNA Methylation
- Gene expression data (miRNA, mRNA, Total RNA, Micro- and Protein-arrays)
- Copy numbers
- https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp



Data archive description

- https://wiki.nci.nih.gov/display/TCGA/Data+archive



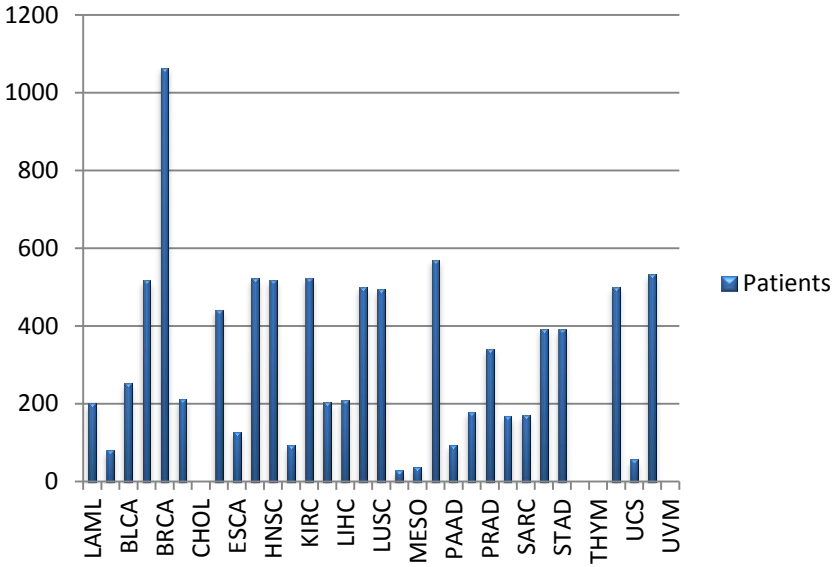
Metadata (Clinical Data)

- Clinical information about the participant
- Information about how participant samples (biospecimens) were processed by the TCGA Biospecimen Core Resource Center (BCR)
- Available in XML and as a flatfile biotab format
- Example of TCGA clinical data:
  - vital status at time of report disease-specific diagnostic information, and initial treatment regimens
  - additional clinical follow up information for some or all participants
- Data elements and fields must follow the rules of XML Schema Document (XSD) http://tcga-data.nci.nih.gov/docs/xsd/BCR/

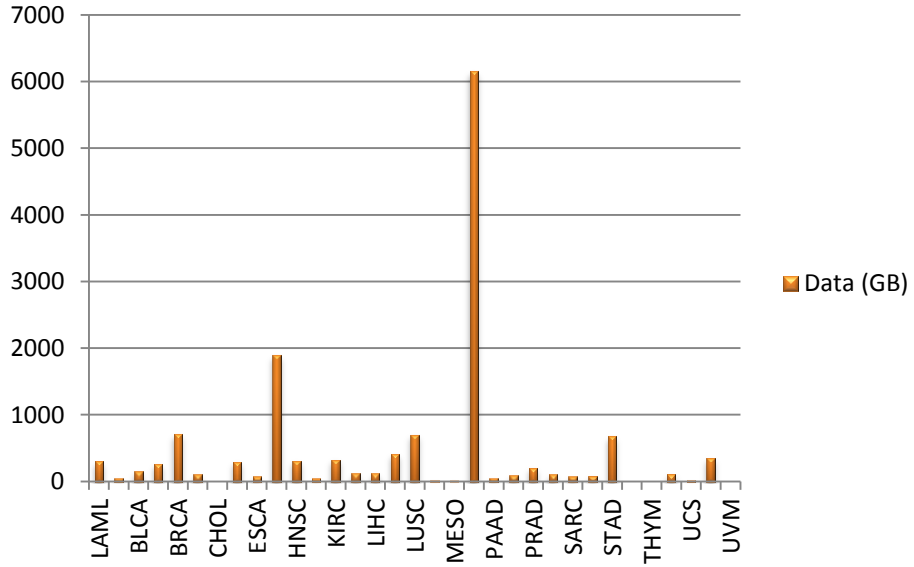
TCGA: some statistics

- 30 different tumor types
- 9404 patients
- 13.45 TB experimental data (clinical and genomic)
- 300 different metadata attributes
  - clinical reports (pdf), images (dcm), blood tests (biotab)

Number of patients



Data quantity (GB)



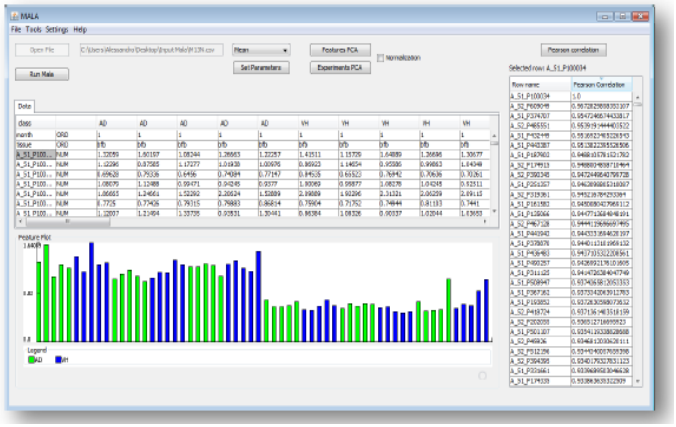
RNA-Seq, DNA-Methylation, Mutations in Cancer

1. Data extraction from The Cancer Genome Atlas (TCGA), the major repository for cancer data
2. Focus on breast cancer (1000 patients)
3. Transformation in GenData format
4. Matrix extraction in GMQL

Patient	$Feature_1$	$\cdots$	$Feature_m$	Class
$sample_1$	$value_{(1,1)}$	$\cdots$	$value_{(1,m)}$	BRC
$sample_2$	$value_{(2,1)}$	$\cdots$	$value_{(2,m)}$	BRC
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$sample_n$	$value_{(n,1)}$	$\cdots$	$value_{(n,m)}$	Control

5. Knowledge extraction with GELA and supervised machine learning techniques

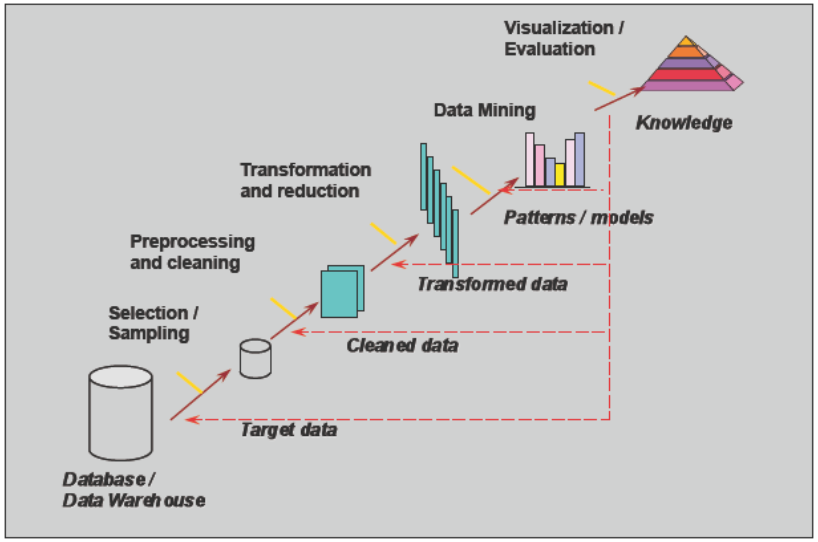
In collaboration with:
Masseroli M., Pinoli P., Fiscon G., Cumbo F., Cestarelli V.



GELA (Gene Expression Logic Analyzer)

Data Mining

- The interdisciplinary field of data mining, which guides the automated knowledge discovery process, is a natural way to approach the complex task of biological data analysis… with the aid of a good bioinformatician
- Data Mining: extraction of knowledge with computer science algorithms for discovering hidden information in heterogeneous data
- The knowledge discovery process:



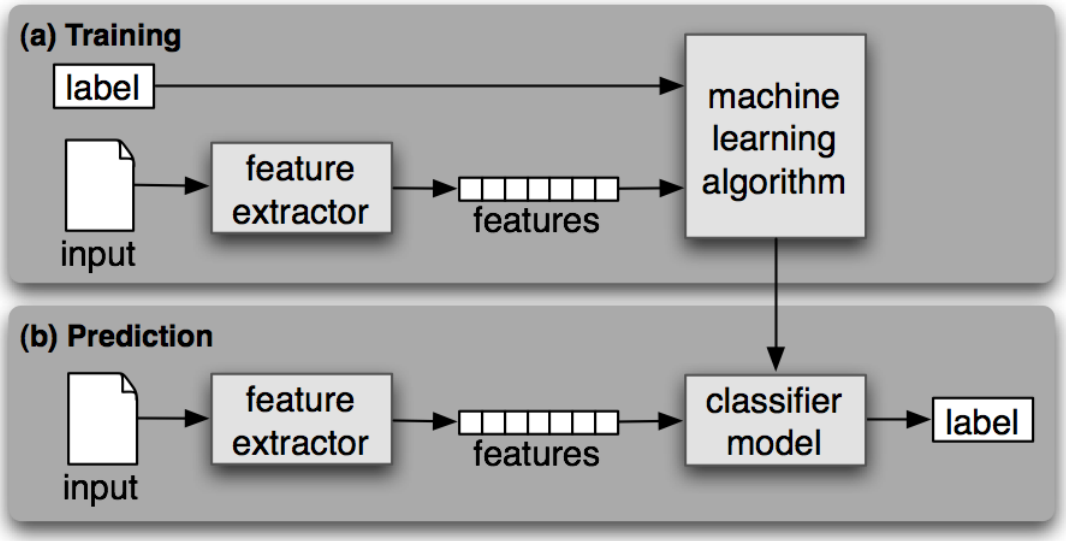
- Main Applications:
  - **Classification**
  - **Clustering**
- Classification is the action of assigning an unknown object into a predefined class after examining its characteristics
- Clustering partitions objects into groups, such that similar or each other related objects are in same clusters

Classification and supervised machine learning

- Goal: assign an unknown sample to a known class starting from its attributes
- The classification problem may be formulated in the following way:
  - given a reference library (training set) composed of samples of known class and
  - a collection of unknown samples (query set or test set)
  - recognize the latter into the class that are present in the library
  - to obtain reliable results
    - the query set has to contain only samples from the same class that are present in the reference library
    - the reference set has to contain a sufficient number of samples for each class (at least 4 samples per class)

Classification and supervised machine learning

- The user has to provide as input a training set (*reference library*) containing samples with a priori known class membership
- Based on this training set, the software computes the classification model
- Subsequently, the classification model can be applied to a test set (*query set*) which contains samples that require classification
- The test set can contain query samples with unknown species membership or, alternatively, samples that also have a priori known species membership, allowing verification of the classifications



Supervised machine learning: Weka classification software

- WEKA (Waikato Environment for Knowledge Analysis) machine learning software is widely adopted for classification
- WEKA contains several methods to perform supervised classification of general problems
- Input a *reference library* in arff format
  - Data and sequences have to be converted
  - sequences have to be of the same region or pre-aligned to the same region

- Weka computes the classification model
- The classification model can be applied to a *query set*
- For using Weka in biomedical data classification reference and query set have to be converted in arff format or csv

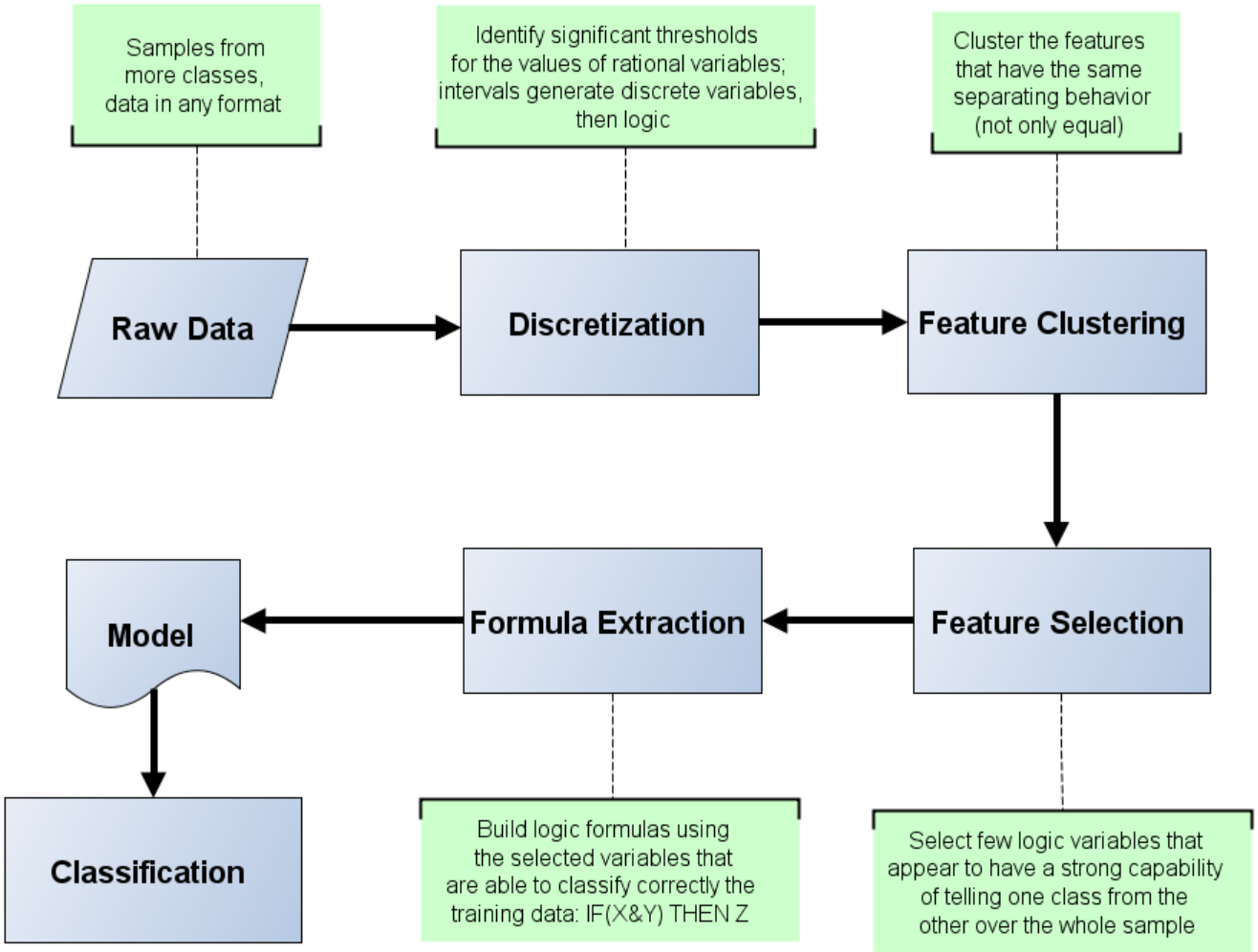
Logic Data Mining

- Logic data mining: classification with logic formulas
- The classifier uses logic propositional formulas in disjunctive or conjuctive normal form (”if then rules”) for classifying the given records, this classification method is also called ruled based
- A logic classifier is a technique for classifying records using a collection of logic propositional formulas (in CNF or DNF), “if… then rules”:
  - Antecedent → Consequent
  - (Condition$_1$) or (Condition$_2$) or … or (Condition$_n$) → Class
  - Condition$_i$: (A_1 op v_1) and (A_2 op v_2) and … and (A_m op v_m)
  - A: attribute
  - v: value
  - op: operator {=, ≠, <, >, ≤, ≥}

If pos13 = A and pos400 = G then the virus is RhinoA

DMB: a logic data mining system

DMB is based on several computational steps, which have been integrated in the software:
1) Discretization 2) Feature clustering 3) Feature selection 4) Formulas extraction
5) Classification and noise reduction 6) Supplementary analysis



Discretization

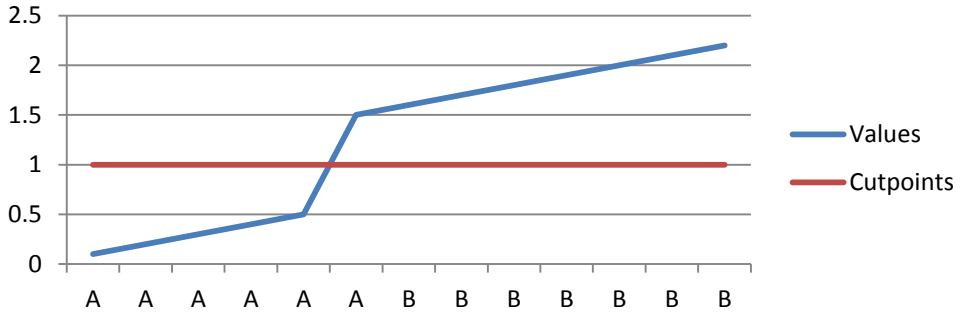
- Logic methods can not deal with raw or numeric data -> transformation of the data
- Ordinal features trivial: eg. A C G T -> 1 2 3 4

Numeric attributes:

- Input: matrix n x m of reals (n rows: samples, m colums: features or attributes)
- For each numeric feature f of the data set
  - Identify a set of intervals of values - Compute the number of samples in each interval
  - Eliminate the empty intervals -Unify the contiguous intervals with the same level of class entropy

The Starting Point

- The numeric real data samples of the feature are sorted in ascending order
- A cut-point is set when a class change occurs between two ordered samples
- The cut-point is defined on the mean value of the two samples that belong to the two different classes

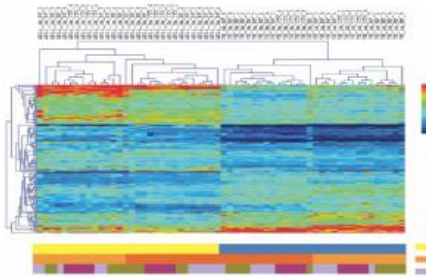


Feat/class	A	A	B	B
A51xxx	0.4	0.5	1.5	1.6
Cutpoint	1.0			
A51xxx_disc	0	0	1	1

Balancing the Penalty Score (PS)

Feature clustering

- Large number of features in biological data sets
- Methods to extract a subset of features able to characterize the classification model
- Aims of feature clustering:
  - To group together similar features
  - To reduce the data to be analyzed
- Discrete Cluster Analysis (DCA):
  - Discretization is applied to numeric features
  - After the discretization step an integer mapping is computed for every feature, that represents the interval in which an experiment falls
  - This integer mapping can be represented in a binary form
  - Two or more features are merged into the same cluster when their binary representation over the intervals is equal (or at a given distance)
  - Features with the same (or similar) discretized profile over the samples are clustered
  - Finally, a feature for each cluster is elected as its representative
  - Clusters composed of a single feature may also be present and are considered as non clustered features



Feature selection

- Feature selection is the identification of a small subset of important attributes or features in a large data set
- Feature selection as a combinatorial problem: Variation of the Set Covering formulation
- The purpose of this formulation is to find a set of given dimensions (beta) that maximises the lower bound in the "separating" information

$$\max \alpha$$

$$\sum_k d_{ij}^k x_k \geq \alpha, \quad \forall i, j, i \neq j$$

$$\sum_k x_k \leq \beta$$

$$x_k \in \{0,1\}$$

$$a_{ij} \in \{0,1\} \Rightarrow d_{ij}^k = \begin{cases} 1 & if\,(a_{ik} \neq a_{jk}) \\ 0 & \text{otherwise} \end{cases}$$

- Data Matrix A (exp. X features)
- $x_i = 1$ if f_i is chosen and 0 otherwise
- each constraint is associated with a pair of items belonging to different classes
- α is a variable that measures the degree of information redundancy

- FS is NP-hard -> At optimality: with contained dimensions; else heuristics…
- GRASP: Greedy Randomized Adaptive Search Procedure, successfully applied to find approx solutions to hard combinatorial problems

Formula Extraction

- After the output of the candidate (cluster of) features (FS) DMB extracts the logic classification formulas
- The Lsquare method is part of DMB for computing the data model, e.g. the separating ”if-then” formulas or rules
- The aim is to extract logic relations that explain the data
- Lsquare approaches this challenge as a sequence of Minimum Cost Satisfiability Problems (MinSat), a combinatorial NPHard optimization problem
- The problem is solved with an algorithm based on decomposition techniques
- DMB computes for every class of the experimental samples the logic classification formulas in Disjunctive Normal Form
- Inequalities in the form of, e.g. ”IF Aph1b<0.47”, and are conjucted in ”AND” and ”OR” clauses

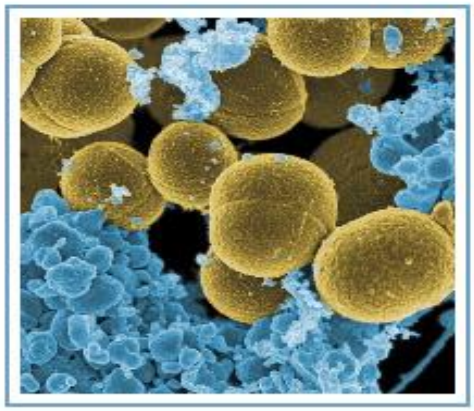
	v_1	v_2	v_3	v_4
A	T	T	F	?
A	T	F	F	T
A	T	F	F	F
B	T	T	T	?
B	F	?	F	T

Classification and noise reduction

- A special classification procedure, which provides a weighting of the logic formulas with the Laplace score and a noise reduction procedure has been developed and integrated in DMB
- Classification:
  - The formulas are firstly weighted with the Laplace Score on the training set and then applied on the test set for performing the classification assignments
  - Additional cut offs of logic formulas with sub-optimal coverage are done by considering the false positive and true positive rates
- The noise reduction applies into three steps of the DMB system:
  - At the discretization level, where it introduces a special purity measure of the intervals
  - At the feature selection level, where it forces the feature selection to the minimal set of variables able to cover the distinct classes of the data set
  - At the formula extraction level, where a pruning of the logic formulas is performed according to two statistical measures of the clauses

DMB (Data Mining Big)

DMB (Data Mining Big) is a set of data analysis tools.
DMB contains a collection of software tools that perform knowledge extraction from data. The methods adopted are based on several models and algorithms that have been developed by a team of researchers, most of them members of the computational and system biology research group. The description of the methods is available in different papers listed in the publication page, while the code can be executed remotely from this website on the servers of IASI - CNR - a research institute of the Italian Research Council. Results are sent by email or visualized on the web interface.



The algorithms used for the DMB System are highly accurated, efficient and innovative. Every single data is checked and examinated accurately by our analysis programs to guarantee an adeguate classification.

The main characteristic of the DMB system is that it extracts knowledge in form of logic rules. DMB takes an input a matrix with the elements on the rows and the attributes on the columns, plus a class label for each element and related labels. Regardless of the form of the input data - rational or discrete numbers, qualitative attributes, binary or logic values, it always returns as output an explanation of the type `if (X`

Latest News

NEW DISCRETIZATION PROCEDURE
November 2010

A new discretization procedure for numeric data sets has been implemented. Check it out.

GRASP 3 IS READY
September 2010

The new feature selection algorithm is released.

NEW CLASSIFICATION ALGORITHM
September 2010

Applications to biological data sets

Species classification with DNA Barcode sequences (BLOG)

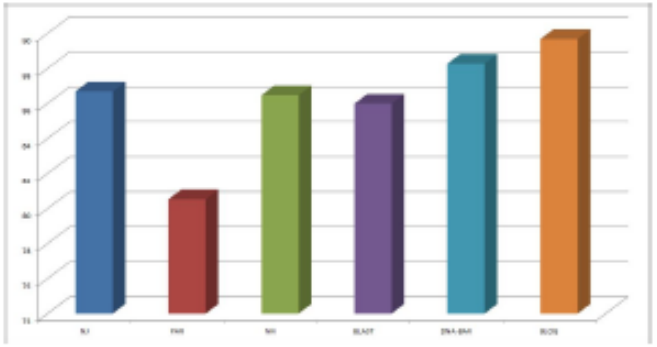


Figure 3: BLOG comparative results (correct percentage rates)

DNA Barcoding of recently diverged species: Relative Performance of Matching Methods. R. Van Velzen, E. Weitschek, G. Felici and F.T.Bakker. *Plos One 7(1):e30490, 2012*

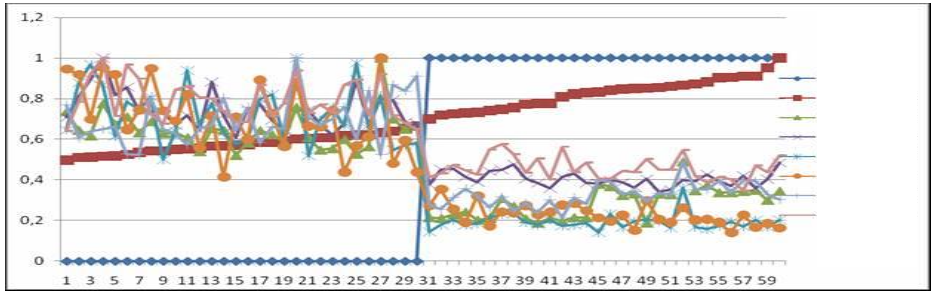
Polyomaviruses identification (DMiB)

```
  1 ---------- --TTATCT-- -GAAGAG--- AAGCAGCTA- --CAA-----
                    ----->
 51 -TGTTTGGAC ---AGT---- ---------- ---------- ---AGACAA-
101 --TTAGTAAA G--------- TAT---CCA- -----GGA-- ----------
151 AGC---AAGC TT-------- ---------- ---------G ACTCA---AG
201 CAGTGTACAT GATTTA---- ---------- -AATATA--- CAA---TCT-
251 -----CCT-- -TAC---ACA CCT------- --GAG----- -TCA---TTC
301 AACACC---- --------GA GTTA------ ---GAATCC- --CCATCT--
351 -CCAAAAAGA AGTGCACCA- --GAGGAGCC TAGCTGTTCT CAGGCAACCC
                                          <----
401 CTCCTAAGAA AAAACATGCA TTTGATGCTT CTTTAGAATT TCCTAAAGAG
        ----->
451 TTGTTAGAGT TTGTTTCACA TGCTGTATTT AGTAATAAGT GTATAACGTG
              <-------
501 C---GTAGTA CAT---ACTA GAGAAAAA-- -GAAGTACTT TAT---AAGT
                         ---->
...
```

Human polyomaviruses identification by logic mining techniques. E.Weitschek, A. Lopresti, G. Felici, G. Drovandi, M. Ciccozzi, M. Ciotti and P. Bertolazzi. *Virology Journal 58(9), 2012*

Applications to biological data sets

Gene expression profiles analysis (MALA)



FONDAZIONE EBRI
"RITA LEVI-MONTALCINI"

If gene Nudt19>0.76 then the individual is healthy

Cluster size	Frequencies
1	3068
2	289
3	86
4	51
5	26
6	14
7	20
8	8
9	6
...	...
244	1
299	1
420	1
441	1
6650	1

Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection. I. Arisi, M. D'Onofrio, R. Brandi, A. Felsani, S. Capsoni, G. Drovandi, G. Felici, E. Weitschek, P. Bertolazzi and A. Cattaneo. *Journal of Alzheimer's Disease 24(4): 72138, 2011*

Clinical patients trials classification (DMiB)

FONDAZIONE EBRI
"RITA LEVI-MONTALCINI"

New diagnostic model for the early diagnosis of Alzheimer's Disease and other dementias, based on Logic Mining of clinical variables). Arisi, Weitschek, et al.;
10th International Conference on Alzheimer's & Parkinson's Diseases, 2011

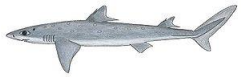
Variables that discriminate between one class form the others			
Normal	**MCI**	**Dementia**	**Depression**
Albumin	Albumin	Albumin	Babcock_1
Anxiety_symptoms	Anxiety_symptoms	Babcock_1	CIRS_cognitive_psychiatric
Azotemia	Azotemia	CIRS_cognitive_psychiatric	CIRS_hypertension_artery
Babcock_1	Babcock_1	CIRS_hypertension_artery	CIRS_skeletal_muscle
CIRS_cognitive_psychiatric	CIRS_cognitive_psychiatric	CIRS_skeletal_muscle	Copy_drawing_corrected
CIRS_hypertension_artery	CIRS_hypertension_artery	Copy_drawing_corrected	Delayed_Recall_corrected
Copy_drawing_corrected	Copy_drawing_corrected		Diffused_hypodensity_CT
Cortical_CT	Cortical_CT	Frontal_hom_hypodensity_CT	FAS
Delayed_Recall_corrected	Delayed_Recall_corrected	How_lives_alone_not	How_lives_alone_not
ECG_patological	ECG_patological	How_long_Stop_drink	How_long_drink
How_lives_alone_not	How_lives_alone_not	IADL_Total	IADL_Total
IADL_Total	How_long_drink	MMSE_Total	Main_job
Main_job	IADL_Total	NPI_Depression	Marital_status
NPI_Depression	Main_job	Years_education	MMSE_Total
Token_test	NPI_Depression		NPI_Depression
	Token_test		Son_daughter_living
			Token_test

DNA Barcode data analysis

- DNA is used for classifying species
- Barcode: small sequence of DNA containing all the information necessary to identify a species of an individual (classification), or to identify new species:
  - Gene region COI for animals (approx. 650 bp)
  - Gene regions rbcl and matk for plants (approx. 1000 bp)
  - Gene regions ITS for fungi (approx 400 bp)
- The sequences are aligned and a positional analysis is performed for distinguishing the different species

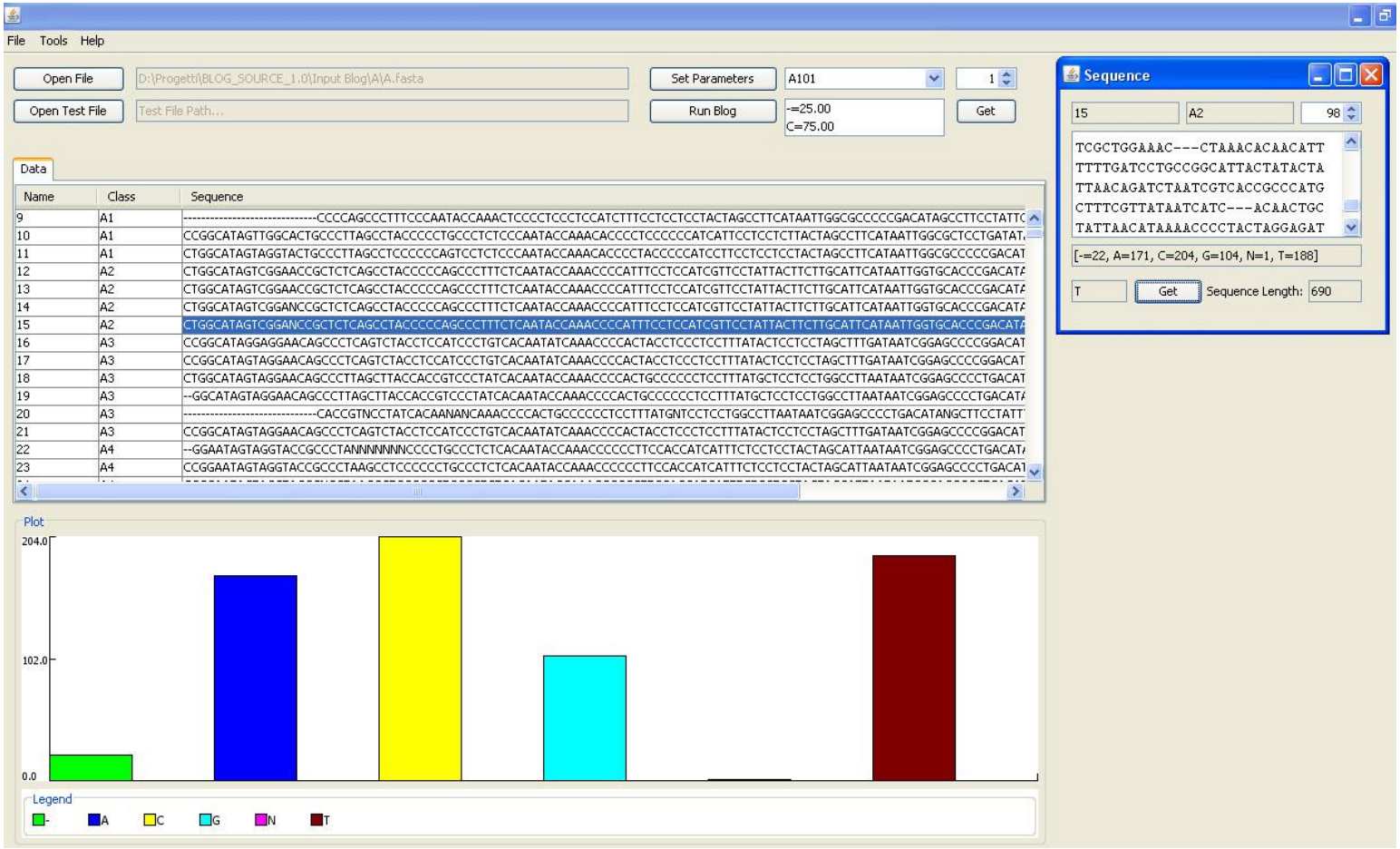
```
IF BASE IN POSITION 466 IS C AND
   BASE IN POSITION 595 IS T THEN SPECIES IS …1

IF BASE IN POSITION 340 IS G AND
   BASE IN POSITION 451 IS A AND
   BASE IN POSITION 493 IS C THEN SPECIES IS … 2

IF BASE IN POSITION 340 IS T AND
   BASE IN POSITION 466 IS A AND
   BASE IN POSITION 625 IS G THEN SPECIES IS … 3
```

If pos3 = A and pos458 = C then the specimen is a



Learning to classify species with barcodes.
P. Bertolazzi, G. Felici and E. Weitschek.
BMC Bioinformatics 10(S-14):7, 2009

BLOG 2.0

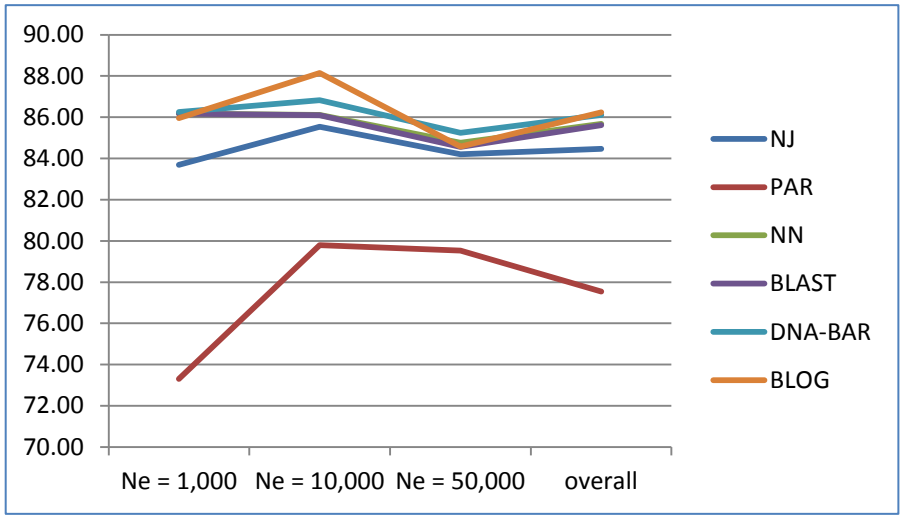


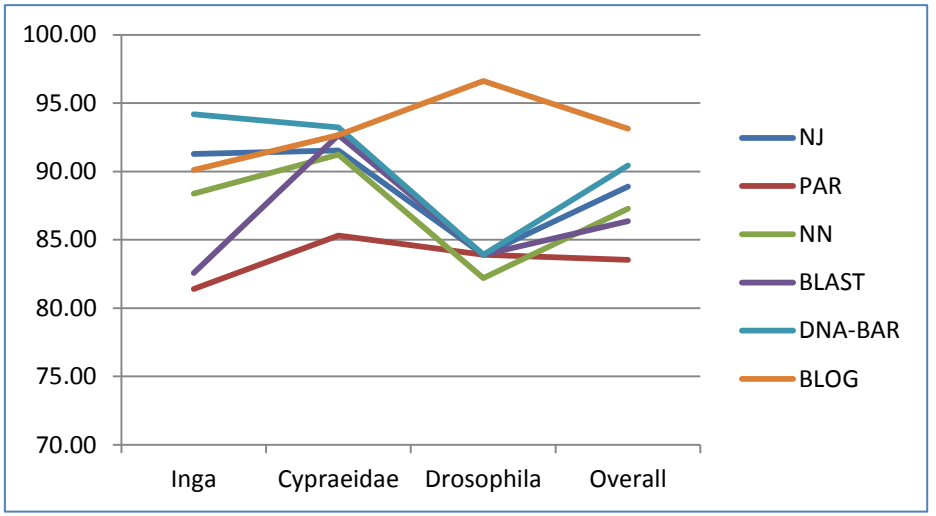
BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it.

E. Weitschek, R. van Velzen, G. Felici and P. Bertolazzi.

Molecular Ecology Resources 2013 (doi: 10.1111/1755-0998.12073)

BLOG vs other methods





Data set	NJ	PAR	NN	BLAST	BAR	BLOG
1,000	83.69	73.31	86.18	86.18	**86.25**	85.96
10,000	85.53	79.79	86.11	86.09	86.83	**88.15**
50,000	84.20	79.53	84.76	84.56	**85.24**	84.58
Overall	84.47	77.54	85.68	85.61	86.11	**86.23**

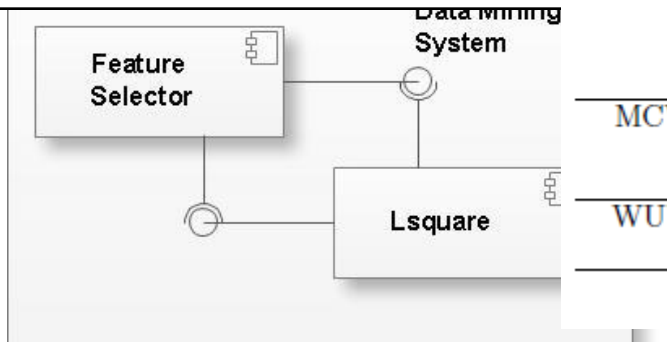
Data set	NJ	PAR	NN	BLAST	BAR	BLOG
Drosophila	83.90	83.90	82.20	83.90	83.90	**96.61**
Inga	91.28	81.40	88.37	82.56	**94.19**	90.12
Cypraeidae	91.53	85.31	91.24	92.66	**93.22**	92.66
Overall	88.90	83.53	87.27	86.37	90.43	**93.13**

DNA barcoding of recently diverged species: Relative Performance of Matching Methods. R. Van Velzen, E. Weitschek, G. Felici and F.T.Bakker. *Plos One 7(1):e30490, 2012*

Polyomaviruses identification

- Aim of this study :
  - find nucleotide positions that allow to distinguish the five human

```
  1 ---------- --TTATCT-- -GAAGAG--- AAGCAGCTA- --CAA-----
                  ----->
 51 -TGTTTGGAC ---AGT---- ---------- ---------- ---AGACAA-
101 --TTAGTAAA G--------- TAT---CCA- -----GGA-- ----------
151 AGC---AAGC TT-------- ---------- ---------G ACTCA---AG
201 CAGTGTACAT GATTTA---- ---------- -AATATA--- CAA-
251 -----CCT-- -TAC---ACA CCT------- --GAG----- -TCA
301 AACACC---- --------GA GTTA------ ---GAATCC- --CC
351 -CCAAAAAGA AGTGCACCA- --GAGGAGCC TAGCTGTTCT CAGG
                                        <----
401 CTCCTAAGAA AAAACATGCA TTTGATGCTT CTTTAGAATT TCCT
          ----->
451 TTGTTAGAGT TTGTTTCACA TGCTGTATTT AGTAATAAGT GTAT
             <-------
501 C---GTAGTA CAT---ACTA GAGAAAAA-- -GAAGTACTT TAT-
                           ---->
...
```



Species	Genes	Formulas
BK	VP1,VP2,VP3	(pos437=A) AND (pos486=C)
JCV	VP1,VP2,VP3	not(pos338=C) AND (pos532=C)
KIV	ST, LT,VP1,VP2,VP3	not(pos294=T) AND not(pos358=T) AND not(pos521=T) AND not(pos532=G)
MCV	ST,LT	(pos199=A) AND not(pos286=T)
WUV	ST, LT, VP1, VP2	not(pos286=T) AND pos425=A AND not(pos474=G)

Table 3.8: Logic formulas for virus classification

Human polyomaviruses identification by logic mining techniques. E.Weitschek, A. Lopresti, G. Felici, G. Drovandi, M. Ciccozzi, M. Ciotti and P. Bertolazzi. *Virology Journal 58(9), 2012*

Alignment free methods for classification

- When analyzing biological sequences with a classical logic data mining approach an overlapping gene region is necessary, due to the fact that an analysis of the characteristics nucleotides present in a determined position for every class is performed

```
IF BASE IN POSITION 466 IS C AND
   BASE IN POSITION 595 IS T THEN SPECIES IS …1

IF BASE IN POSITION 340 IS G AND
   BASE IN POSITION 451 IS A AND
   BASE IN POSITION 493 IS C THEN SPECIES IS … 2

IF BASE IN POSITION 340 IS T AND
   BASE IN POSITION 466 IS A AND
   BASE IN POSITION 625 IS G THEN SPECIES IS … 3
```

```
Species   123456789
Species1  ACGTTAACA
Species1  GTCAGTCCA
Species2  GTCGTATAT
Species2  AGCTACGAT
```

- Often this is obtained through an alignment between the sequences to analyze, i.e. align areas of the sequences sharing common properties
- Optimal methods for sequence alignments rely on dynamic programming
- Limits of alignment:
  - high computational requirements (exponential in the length of the sequences)
  - non consideration of recombinations and shuffling of the sequences
  - often the sequences are not alignable, e.g. non coding regions, or very hard to align, e.g. whole genomes

R.A.2012: Alignment free methods for classification

- Solutions: Alignment free methods
  - Basic Ingredient: Similarity/Distance functions between strings
  - Similarity of two strings is assessed based only on the DICTIONARY of substrings that appear in the strings, irrespective of their relative position
  - Advantages: Speed and Scalability; Time linear in the size of the input
- Alignment free k-mer frequency count analysis:
  - oligomers frequencies: computation of the substrings frequencies of a given length k
  - the k-mers are computed by counting the occurrences of the substrings with a sliding window of length k, starting at position 1 and ending at position n−k+1
  - frequency vector, where each component of the vector is associated to the frequency of a particular k-mer
  - coordinate space, that is mathematically tractable
- Combination of alignment free methods and logic data mining:

	Seq1	Seq2	...	Seq3	Seq4	...
	Vertebrate	Vertebrate	...	Invertebrate	Invertebrate	...
AAA	0.46	0.26	...	0.24	0.26	...
AAC	0.12	0.16	...	0.23	0.24	...
AAG	0.13	0.23	...	0.23	0.22	...
...	...	...	...	...	...	...

  - k-mer frequency matrix as input to a logic classifier and extract the classification formulas

R.A.2012: Alignment free and logic mining results

- Whole genome bacteria classification:
  - Whole genomes are very difficult to align, because of their length
  - The problem of multiple alignment is hard and grows exponentially in the size of the input (length and number of sequences)
  - The method is applied for classifying bacteria in the different levels of the phylogenetic tree (phylum, class, family, genus, species) by analyzing their whole genome
  - 1964 bacteria whole genome sequences downloaded from GenBank

Level	JRip	Ridor	Part	DMB	Average
Species	98.14	97.21	98.71	98.90	98.24
Genus	88.77	86.34	85.53	85.84	86.62
Order	78.44	75.28	76.32	76.08	76.53
Class	80.81	79.86	81.08	80.72	80.61
Phylum	83.80	80.99	80.59	81.54	81.73
Average	85.99	83.93	84.44	84.61	84.74

UNIVERSITÀ DEGLI STUDI DI PADOVA

  - Very promising results, especially at the species level (98% correct classification rates)

k	correct-rate-cv mean	std-devcorrect-rate-cv
3	95.60483333	3.349375967
4	97.66783333	0.744334435
5	97.94316667	0.600806291
6	97.24783333	0.822828736
7	96.34316667	0.882061078
mean	96.96136667	1.86129509

R.A.2012: Alignment free and logic mining results

- Conserved non coding sequences (CNEs):
  - CNEs are not able to be aligned, because of their noncoding nature
  - Datasets:
    1) human coding vs human CNEs 2) amniotic vs mammalian
    3) vertebrate CNEs vs invertebrate CNEs
  - AF + Logic data mining

DataSet	JRip	Ridor	Part	DMB	Average
1	91.68	90.55	91.25	91.37	91.21
2	93.89	93.93	91.20	91.46	92.62
3	93.88	93.48	94.87	93.05	93.82
Average	93.15	92.65	92.44	91.96	92.55

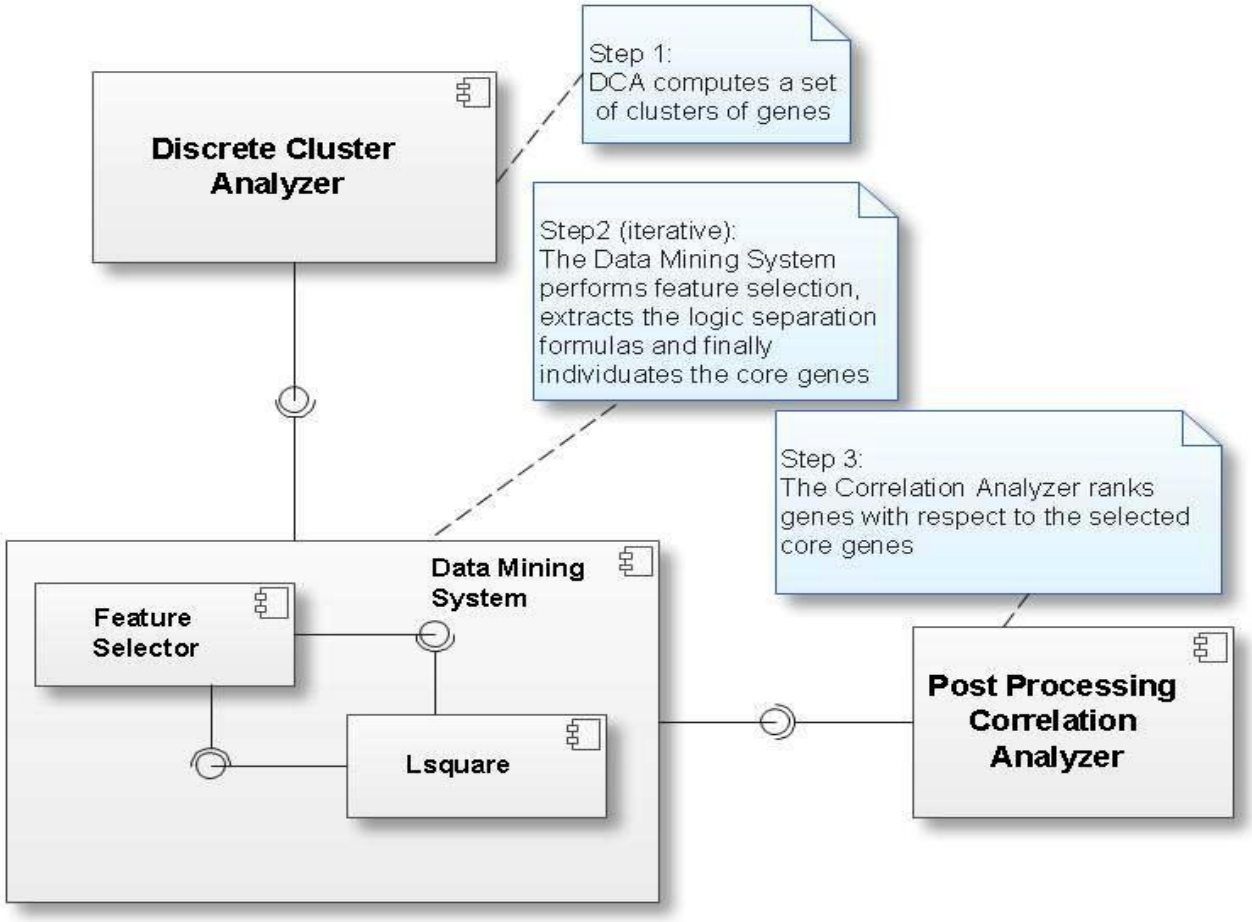
DEMOKRITOS
NATIONAL CENTER FOR SCIENTIFIC RESEARCH

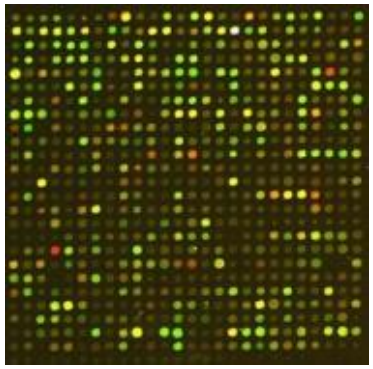
  - GC content analysis [KB95]

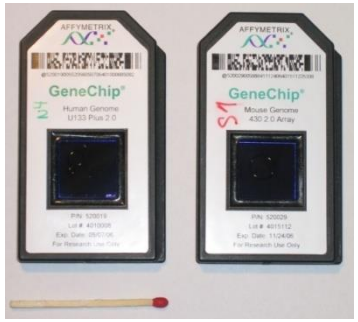
DataSet	JRip	Ridor	Part	DMB	Average
1	65.48	61.38	63.63	64.75	63.81
2	74.79	71.94	73.34	72.68	73.18
3	60.10	60.18	58.99	59.97	59.81
Average	66.79	64.50	65.32	65.80	65.60

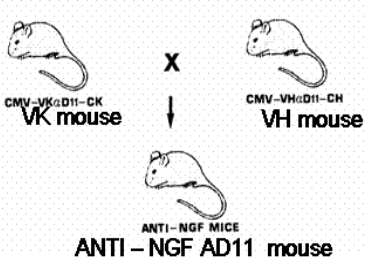
  - ➢ if 719.4 <= freq(CGCG) < 1075.3 OR 719.4 <= freq(CTAG) < 1075.3 then the sequence is coding

Gene expression profiles analysis









e) Post processing and correlation analysis

Gene expression profiles analysis results

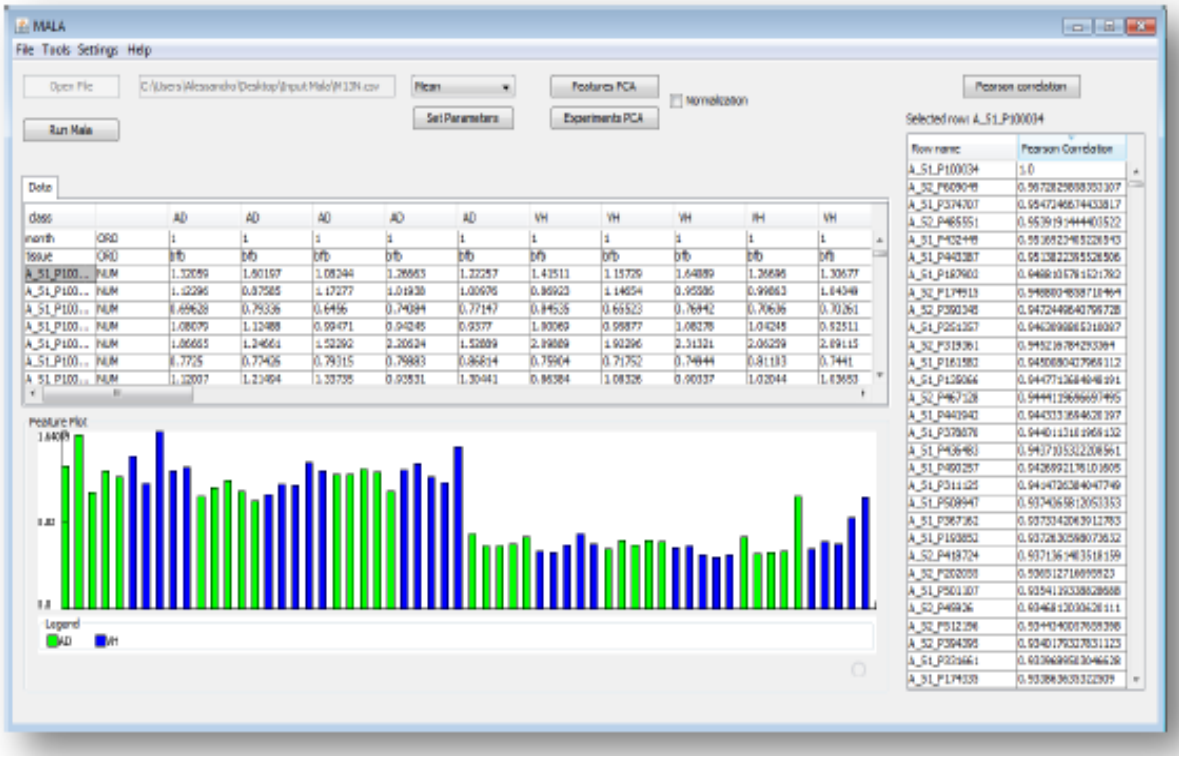
- There are few genes (9) that are able, one by one, to separate exactly all the healthy from the sick mice (iterative application of the method) in leave-1-out cross validation
- More genes are strongly co-regulated with the 9 genes network

Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection.
I. Arisi, M. D'Onofrio, R. Brandi, A. Felsani, S. Capsoni, G. Drovandi, G. Felici, E. Weitschek, P. Bertolazzi and A. Cattaneo. *Journal of Alzheimer's Disease 24(4): 72138, 2011*

Cluster size	Frequencies
1	3068
2	289
3	86
4	51
5	26
6	14
7	20
8	8
9	6
...	...
244	1
299	1
420	1
441	1
6650	1

If gene Nudt19 > 0.76 then the individual is healthy

MALA: MicroArray Logic Analyzer



Aims of MALA:

1) To cluster the microarray gene expression profiles, in order to reduce the amount of data to be analyzed, and to detect similar gene groups

2) To classify the microarray experiments

Computational steps:

5 main computational steps

Software architecture:

Pipes and filters

Releases:

Major operating systems

MALA: A microarray clustering and classification software.
E. Weitschek, G. Felici and P. Bertolazzi.
IEEE DEXA Workshops 2012:201-205

Gene expression profiles analysis results

- Some of the most commonly used classification algorithms were tested on the same Alzheimer data sets

method	settings	early stage	late stage	model
MALA	no settings	100.0	100.0	yes
SVM	polykernel=2	96.66	100.0	no
RF	trees=100	96.66	94.91	no
C4.5	unpruned, minobj=2	98.33	98.30	yes
KNN	k=2	70.00	86.44	no

- Other tests have been performed on data sets downloaded from public repositories Array-Express and GEO

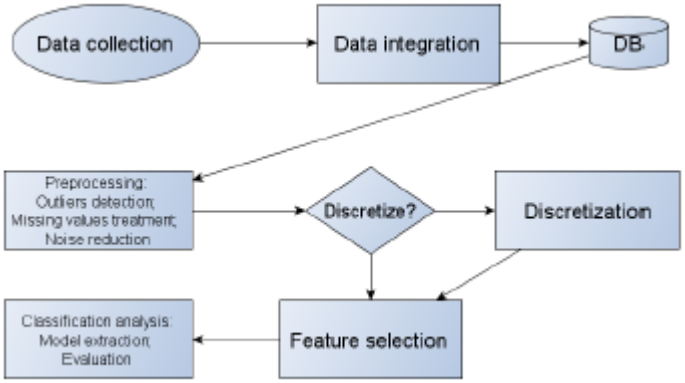
method	settings	MsDiagnostic	Psoriasis	model
MALA	no settings	94.94	100.0	yes
SVM	polykernel=2	90.45	98.86	no
RF	trees=100	91.57	98.86	no
C4.5	unpruned, minobj=2	87.08	97.16	yes
KNN	k=2	87.64	99.43	no

- MALA slightly dominates the other methods and produces a easy human interpretable classification model, i.e. additional knowledge gain for the natural scientist
- The second best is SVM, that unfortunately produces classification models whose interpretation is very difficult for human beings
- Advantages of MALA: to extract meaningful and compact models; clustering capabilities availability as an integrated tool

Clinical Patients

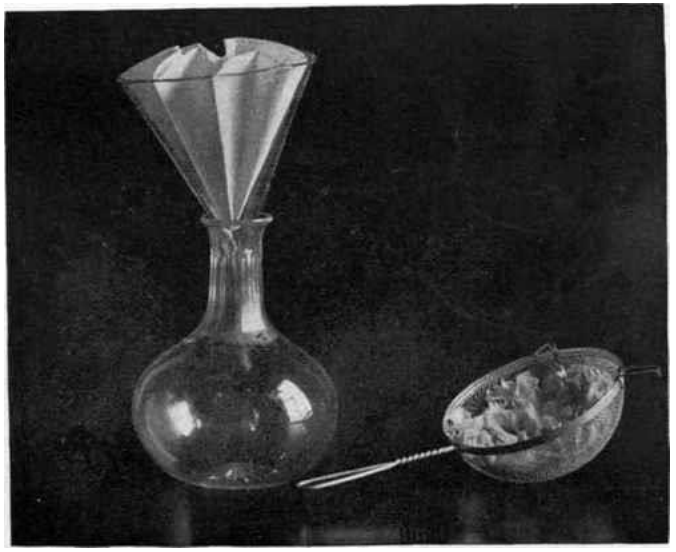
- The aim of the project is to design a new diagnostic model and possibily a work-flow for the early diagnosis of dementias
- Clinical patients trial samples were collected in different Italian health structures
- Collected data included demographic characteristics, medical history, pharmacological treatments, clinical and neurological examination, psychometric tests, laboratory blood tests, imaging…
  (700 features x 5000 experiments)
- The data set:

class	# of patients	# of trials	% missing values
Demented	3110	722	29.45
Depressed	375	722	37.13
MCI	406	722	25.07
Normal	365	722	30.41
Psicotic	49	722	31.77
Uncertain	423	722	31.48
Total	4728	722	30.88



Preprocessing and missing values

- Fundamental in clinical data analysis
- Produces clean data and is crucial for obtaining reliable knowledge by the data mining algorithms
- Perform common statistical analysis for every trial:
  mean, standard deviation, minimum, maximum, number of missing values, modal value, etc.
- Trials filtering:
  - exclusion of the trials with a high number of missing values, e.g. more than 20%
  - filling of missing values. e.g. by the class mean
  - outliers detection by mean - standard deviation - minimum - maximum values analysis and records removal
  - correction of the mixed attributes types



Preprocessing and missing values

- Missing values treatment:
  - Most common attribute value
  - Assigning all possible values of the attribute
  - Ignoring examples with unknown attribute values
  - Event-Covering method
  - Treating missing attribute values as special values
  - Filling of missing values with the class mean (for numeric variables)
  - Filling of missing values with the class modal value (for categorical variables)

v1	v2	v3	v4	v5
.	4		4	4
3	3	6	6	6
7	2	9	9	9
4	.	.	4	4
5	.	.	5	5
.	.	.	0	.
0	3	3	3	3

- Special scripts and programs are necessary

Clinical Patients

- Results:

Variables that discriminate between one class form the others			
Normal	**MCI**	**Dementia**	**Depression**
Albumin	Albumin	Albumin	Babcock_1
Anxiety_symptoms	Anxiety_symptoms	Babcock_1	CIRS_cognitive_psychiatric
Azotemia	Azotemia	CIRS_cognitive_psychiatric	CIRS_hypertension_artery
Babcock_1	Babcock_1	CIRS_hypertension_artery	CIRS_skeletal_muscle
CIRS_cognitive_psychiatric	CIRS_cognitive_psychiatric	CIRS_skeletal_muscle	Copy_drawing_corrected
CIRS_hypertension_artery	CIRS_hypertension_artery	Copy_drawing_corrected	Delayed_Recall_corrected
Copy_drawing_corrected	Copy_drawing_corrected		Diffused_hypodensity_CT
Cortical_CT	Cortical_CT	Frontal_hom_hypodensity_CT	FAS
Delayed_Recall_corrected	Delayed_Recall_corrected	How_lives_alone_not	How_lives_alone_not
ECG_patological	ECG_patological	How_long_Stop_drink	How_long_drink
How_lives_alone_not	How_lives_alone_not	IADL_Total	IADL_Total
IADL_Total	How_long_drink	MMSE_Total	Main_job
Main_job	IADL_Total	NPI_Depression	Marital_status
NPI_Depression	Main_job	Years_education	MMSE_Total
Token_test	NPI_Depression		NPI_Depression
	Token_test		Son_daughter_living
			Token_test

(MMSETotale $\geq$ 24) and (IADLFarmaci1 $\geq$ 1) and (HachinskiScore $\geq$ 4) and (FluiditaVerbaleCategorieCorretto $\leq$ 10) and (CorniFrontaliIpodensitaTAC $\leq$ 0) OR
(MMSETotaleRettificato $\geq$ 22.5) and (MatriciCorretto $\geq$ 34) and (AssurditaTestVerbali $\leq$ 13) and (GB $\geq$ 4.9) OR
(MMSETotaleRettificato $\geq$ 23.200001) and (IADLFarmaci1 $\geq$ 1) and (DatiTestBaseValutazioneDisturbiIDUtente $\leq$ 3) and (regione $\leq$ 6) and (Forgetting $\geq$ 4) OR
(MMSETotaleRettificato $\geq$ 22) and (IADLSpostamenti2 $\geq$ 4) and (DepressionePsichiatrico $\leq$ 0) and (regione $\leq$ 2) and (CIRSComorbilitaComplessa $\geq$ 1) OR
(MMSETotale $\geq$ 23) and (TokenTestCorretto $\geq$ 30) and (RavenTestCorretto $\geq$ 215) and (FamiliaritaDemenza $\leq$ 0) and (RichiamoImmediatoCorretto $\leq$ 34) and (CIRSVascolareLinfatico $\leq$ 1)

- Classification Results

method	settings	correct %	discretization	model
DMB	no settings	87.36	yes	yes
JRIP	opt. run=10	86.86	no	yes
C4.5	unpruned, minobj=2	88.49	no	yes
SVM	polykernel=1	90.313	yes	no
KNN	k=1	78.27	yes	no

Clinical data mining: problems, pitfalls and solutions.
E. Weitschek, G. Felici and P. Bertolazzi.
IEEE DEXA 2013

Noise reduction on clinical patients (TAG SNP)

- Single Nucleotide Polymorphism (SNPs) are positions of the DNA sequences where the differences among individuals are embedded
- Several experiments with the DMB system were performed injecting a controlled amount of noise in simulated ad hoc noisy data sets of SNPs
- In particular, the new noise reduction variants of DMB were tested on the data provided by Thorsten Lehr and described in
- Lehr evaluated the performance of three ruled based classifier algorithms RIPPER, RIDOR and PART and ranked them according to the original model similarity (scale from A to 0)
- A very good behaviour of DMB was achieved: the DMB systems got 12 A, 6 B, 12 C, 6 D and 0 0 scores, on a scale from 0 to 4 the average score is **2.66**
- DMB gets rid of different overfitting effects due to the introduction of noise
- The best performing method got similar results (16 A, 5 B, 4 C, 2 D, 9 0), on a scale from 0 to 5 the average score is **2.47**

EEG signal processing and patients classification

- Electro Encefalo Grams (EEG) of 102 patients:
  - 51 Alzheimer Diseased
  - 37 Mild Cognitive Impairment
  - 14 Control
- 19 electrods
- Signal duration of 300 seconds
- sampling frequency of 1024 or 256 sp/s
- 9 GB of raw signals
- Feature extraction:
  - only 180 seconds for each signal
  - transform each one at 256 sp/s
  - spectral analysis and Fast Fourier Transform

CENTRO NEUROLESI "BONINO-PULEJO" MESSINA

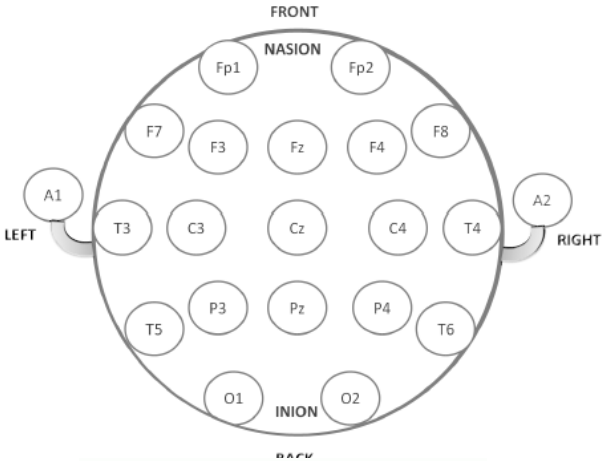
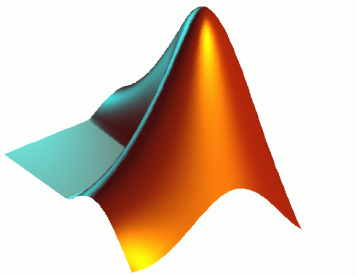


TABLE I
EXAMPLE OF THE EXPERIMENT DATASET WITH P=n^o OF PATIENTS, M=n^o OF ELECTRODES, N=n^o OF FOURIER COEFFICIENTS

Patient	$Fourier_{(1,1)}$	⋯	$Fourier_{(M,MN)}$	Class
$sample_1$	$value_{(1,1)}$	⋯	$value_{(1,MN)}$	AD
$sample_2$	$value_{(2,1)}$	⋯	$value_{(2,MN)}$	MCI
⋯	⋯	⋯	⋯	⋯
$sample_P$	$value_{(P,1)}$	⋯	$value_{(P,MN)}$	Control



EEG signal processing and patients classification

- We apply the FFT functions to the collected data in the following way:
  - taking into account the 180-second signal and extracting N Fourier Coefficients (N equal to 16 and 32);
  - dividing the signal in 6 epochs of 30 seconds and extracting for each one 16(32) Fourier Coefficients

- Classification:

TABLE II
CLASSIFICATION PERFORMANCES [%] FOR THE 180-SECONDS EEG SIGNALS WITH $N = 16$, n^o OF AD(MCI) INSTANCES=63(51), n^o OF AD *vs* MCI INSTANCES=86, L-1-OUT=63,51,86 FOLDS FOR AD *vs* CT, MCI *vs* CT, AD *vs* MCI, RESPECTIVELY

Sampling	AD *vs* CT			MCI *vs* CT			AD *vs* MCI		
	SVM	J48	DMB	SVM	J48	DMB	SVM	J48	DMB
1-1-out	67	60	**70**	61	**84**	65	55	64	52
80-20%	69	54	54	60	60	56	47	**76**	45
training	**90**	**98**	**94**	**90**	**98**	**100**	**87**	**98**	**94**

TABLE III
CLASSIFICATION PERFORMANCES [%] FOR THE 30-SECONDS EPOCHS OF THE EEG SIGNALS WITH $N = 16$, n^o OF AD(MCI) INSTANCES=63(51), n^o OF AD *vs* MCI INSTANCES=86, L-1-OUT=63,51,86 FOLDS FOR AD *vs* CT, MCI *vs* CT, AD *vs* MCI, RESPECTIVELY

Sampling	AD *vs* CT			MCI *vs* CT			AD *vs* MCI		
	SVM	J48	DMB	SVM	J48	DMB	SVM	J48	DMB
1-1-out	54	**79**	62	51	**75**	61	38	**71**	47
80-20%	69	**77**	**70**	60	**80**	55	47	47	56
training	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**98**	**100**

Collaborations

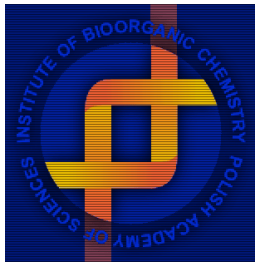
CONSORTIUM FOR THE BARCODE OF LIFE

PAN
POLISH ACADEMY of SCIENCES

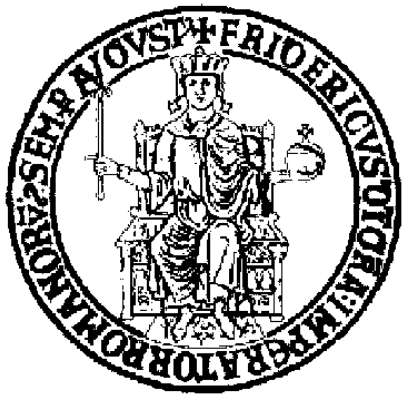
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Boehringer Ingelheim

UNIVERSITÀ DEGLI STUDI DI PADOVA



FONDAZIONE EBRI
"RITA LEVI-MONTALCINI"



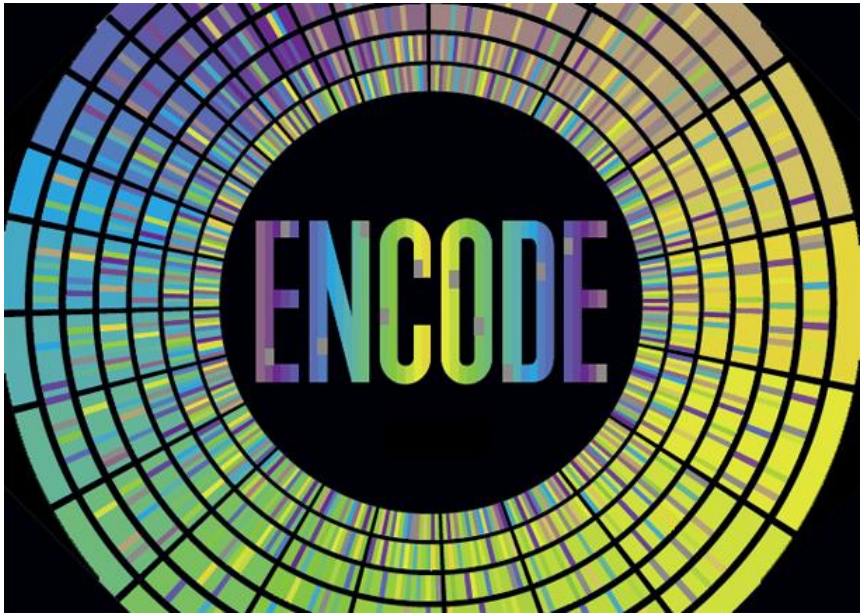
UTD

DEMOKRITOS
NATIONAL CENTER FOR SCIENTIFIC RESEARCH

Project: ENCODE

Encyclopedia of DNA Elements

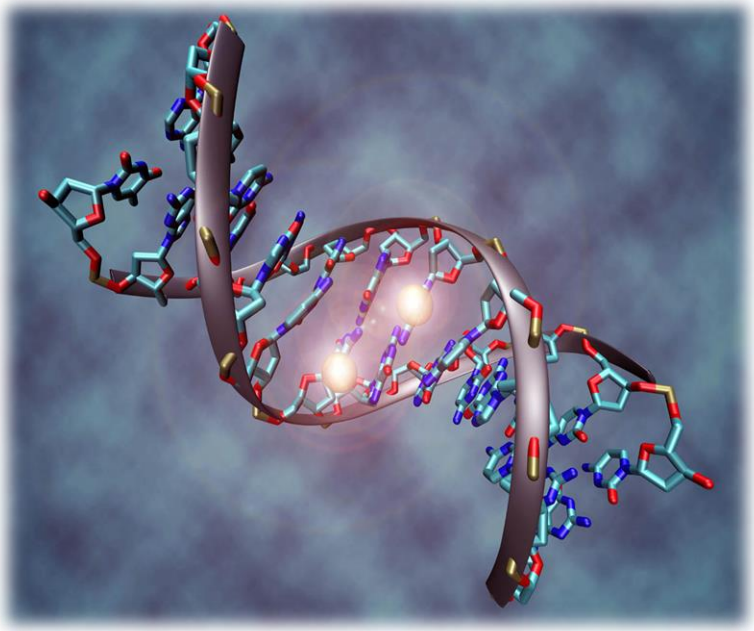
- **Statistical analysis of the ENCODE database:**
  - http://genome.ucsc.edu/ENCODE/
  - ENCODE contains a comprehensive collection of genomes and sequencing experiments
  - Data conversion and analysis with supervised methods
  - Focus on Next Generation Sequencing experiments



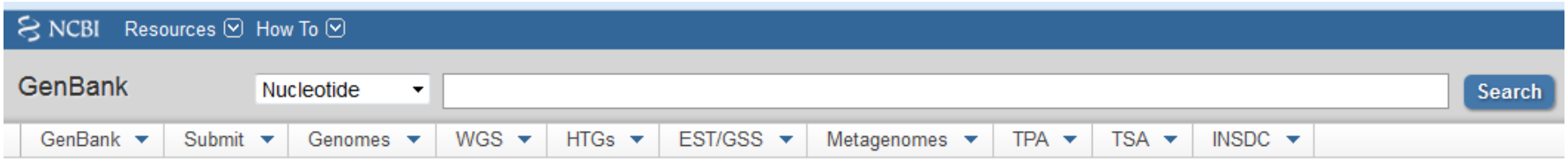
Project: TCGA

THE CANCER GENOME ATLAS

- **Metadata management and analysis in the TCGA database:**
  - http://cancergenome.nih.gov/
  - TCGA contains a comprehensive collection of genomic, clinical and patient data affected by different cancer types
  - Metadata management (XML) and analysis



Project: GenBank

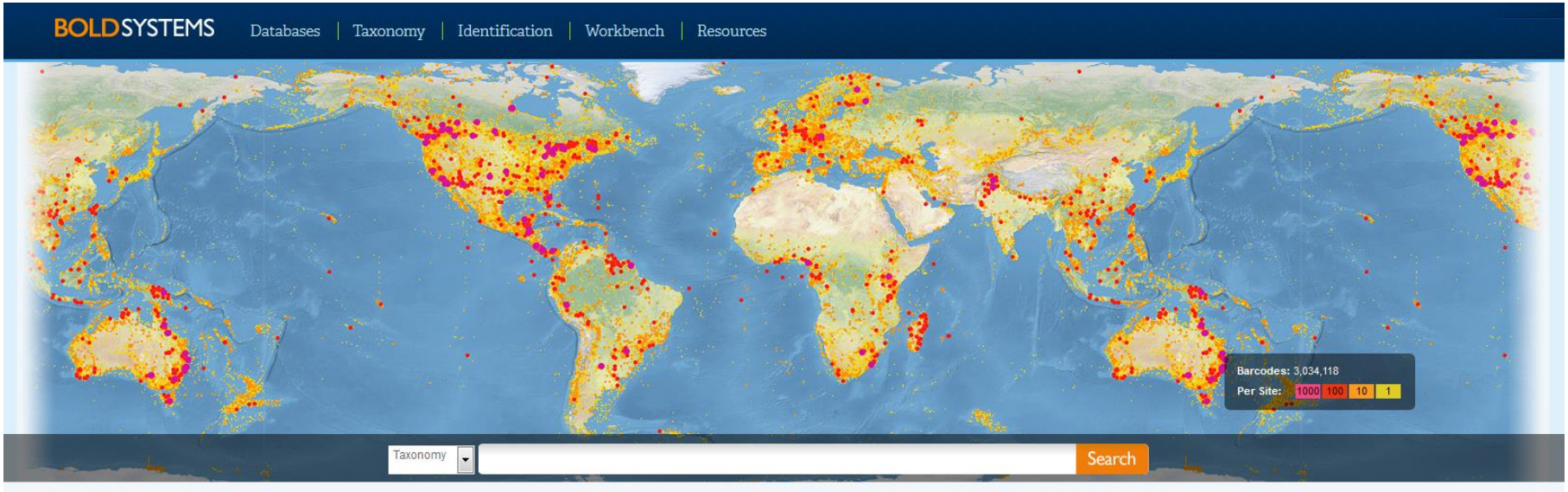


- **Statistical analysis of the GenBank database**
  - http://www.ncbi.nlm.nih.gov/genbank/
  - GENBANK contains a comprehensive collection of public DNA and protein sequences
  - Application of alignment and supervised machine learning algorithms

Project: DNA Barcoding

CONSORTIUM FOR THE BARCODE OF LIFE

- **Statistical analysis of the BOLD Database**
  - http://www.boldsystems.org
  - Bold contains a comprehensive collection of specimen to species assignments through DNA Barcode sequences
  - Application of alignment free techniques and supervised machine learning algorithms
  - Focus on multilocus DNA Barcoding

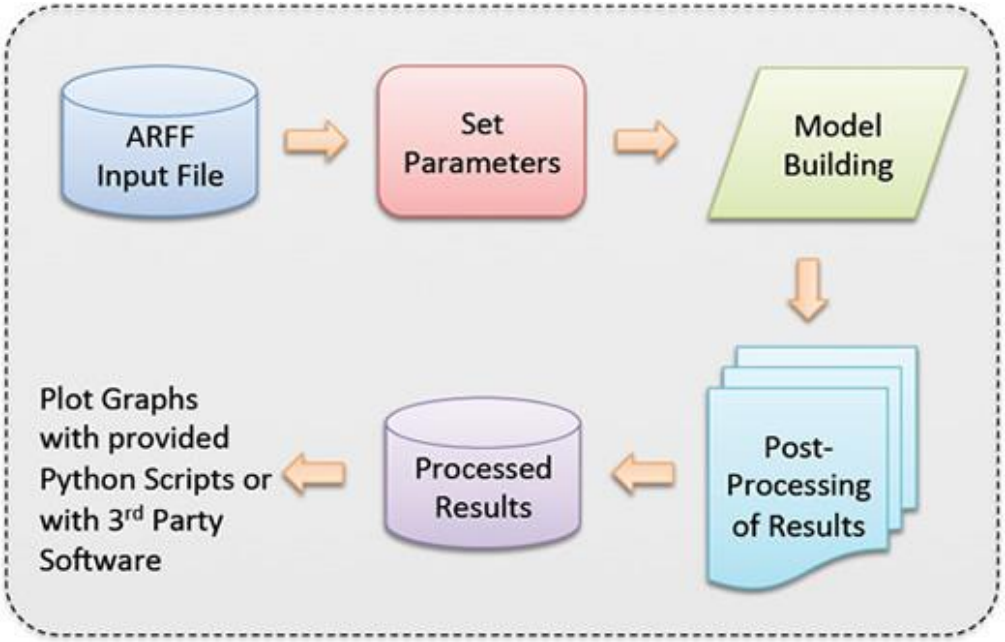


Project: Automatic parameter tuning for data mining

AutoWeka

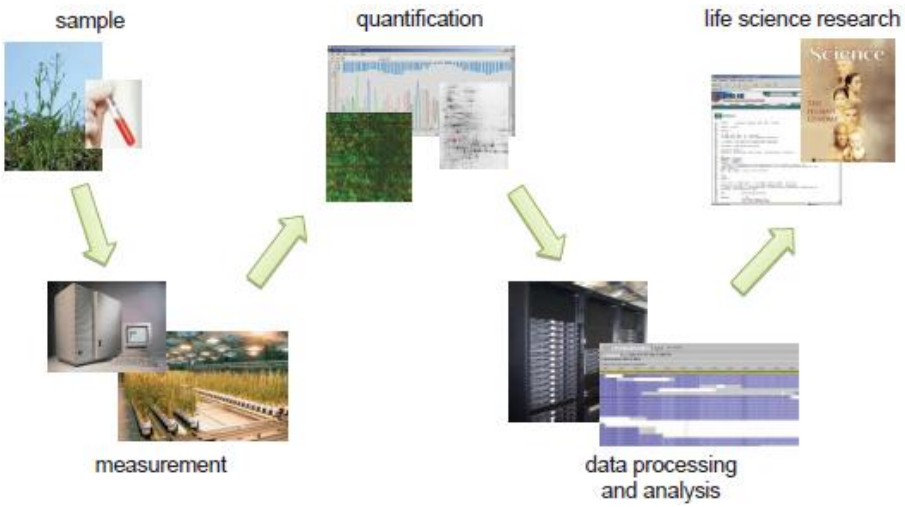
An Automated Data Mining Software Based on Weka

- **Automatic parameter tuning for data mining with AutoWeka**
  - http://www.cs.ubc.ca/labs/beta/Projects/autoweka/
  - AutoWeka considers the problem of simultaneously selecting a learning algorithm and setting its hyperparameters
  - Application of AutoWeka to a big biomedical data set



Project: LIMS

- LIMS: Laboratory Information Management Systems
- Open source LIMS for NGS:
  - GNomEx
  - LABKey Server
  - Galaxy LIMS
  - openBis
  - Wasp
  - Brad Chapman
  - Lims Light System



- Task: Analysis and tests on NGS data of a LIMS

Project: Clinical data cleaning

- **Cleaning a real clinical data set collected from different Italian Health Structures**
  - http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6621352&tag=1
  - Preprocessing and missing values
  - Noise reduction

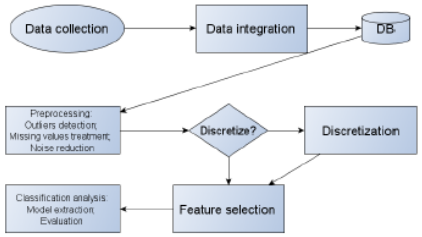
class	# of patients	# of trials	% missing values
Demented	3110	722	29.45
Depressed	375	722	37.13
MCI	406	722	25.07
Normal	365	722	30.41
Psicotic	49	722	31.77
Uncertain	423	722	31.48
Total	4728	722	30.88

Clinical data mining: problems, pitfalls and solutions

Emanuel Weitschek*†, Giovanni Felici† and Paola Bertolazzi†
*Department of Computer Science and Automation
University Roma Tre, Rome, Italy
†Institute of Systems Analysis and Computer Science
National Research Council, Rome, Italy
Email: {emanuel.weitschek, giovanni.felici, paola.bertolazzi}@iasi.cnr.it

Project: Clinical data standardization

- **Study and application of clinical data standards**
  - Issue: when the variables names are different in the diverse data sets
  - This problem may be solved with the creation of ontologies or clinical and laboratory variables names standardization,
  - Logical Observation Identifiers Names and Codes (LOINC): a universal code system for identifying laboratory and clinical observations (http://loinc.org)
  - Health Level Seven International organization (http://www.hl7.org)

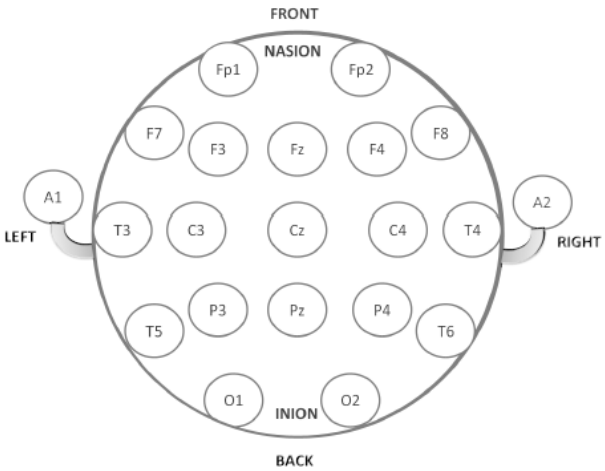


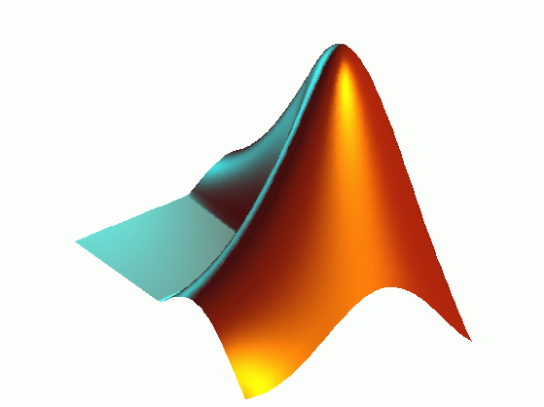
Project: EEG signals preprocessing

EEGLAB

an open source environment for electrophysiological signal processing

- **Study and application of EEG lab for Matlab to Alzheimer patient samples**
  - http://sccn.ucsd.edu/eeglab/
  - Study of the ad-hoc tool EEGlab for MATLAB to perform an effective EEG signal preprocessing
  - Application to a private EEG data set (9GB)





Project: Workflow systems

- **Study and application of a workflow system for analyzing biomedical data**
  - The majority of biomedical data analysis task require the application of different software and tools composed in a pipeline
  - Several tools allow the composition of state of the art bioinformatics programs in a workflow:

Taverna (http://www.taverna.org.uk/)

Galaxy (http://galaxy.psu.edu/)

Anduril (http://csbl.fimm.fi/anduril/site)

Project: MyExperiment

myExperiment makes it easy to **find**, **use** and **share scientific workflows** and other **Research Objects**, and to build **communities**.

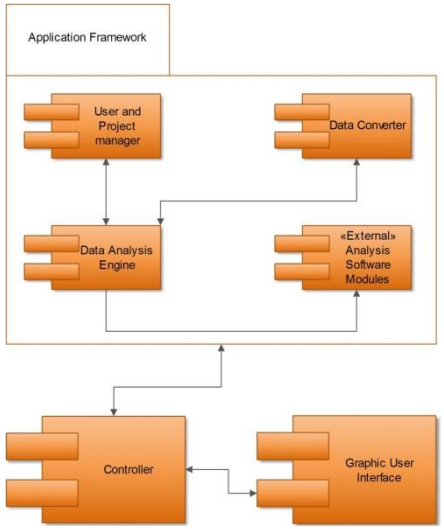
- **Analysis of the MyExperiment platform**
  - http://www.myexperiment.org/
  - myExperiment makes it easy to find, use and share scientific workflows and other Research Objects, and to build communities

Project: CoSys IT infrastructure

CoSys© IT infrastructure:

- Computational Systems Biology Information Technology Infrastructure
- It is the supporting IT facility for the consortium.
- Main feature: data modelling and analysis workflows.

- The CoSys© (Computational Systems Biology Infrastructure) is an integrated web information system
- It allows experimental and project Systems Biology data
  - acquisition
  - storage
  - indexing
  - search
  - analysis
  - visualization
  - Exchange
- **Task: Participation to the definition of workflows and software modules**



Contacts

- **Emanuel Weitschek**
  University Roma Tre
  Department of Computer Science and Automation
  Rome, Italy
  www.dia.uniroma3.it/~emanuel
  emanuel@dia.uniroma3.it

Thanks for your attention!