



"This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement N°603946"



HEALS

Health and Environment-wide Associations based on Large population Surveys



HEALS

Health and Environment-wide Associations
based on Large population Surveys

FP7-ENV-2013- 603946

<http://www.heals-eu.eu/>

7.1 Data infrastructure and data mining model of internal exposome

WP 7 Novel bioinformatics for predictive biomarker discovery

Version 1 (28/04/2015)

Lead beneficiary: AUTH

Date: 28/04/2015

Nature: Report - R

Dissemination level: Public - PU




 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	2/61

TABLE OF CONTENTS

1	SUMMARY	4
2	INTRODUCTION - GENERAL CONSIDERATIONS	6
3	METHODOLOGICAL FRAMEWORK FOR BIG DATA ANALYTICS	8
3.1	Omics workflow	8
3.2	Coupling <i>in vitro</i> and <i>in silico</i> analysis	10
4	STATISTICAL METHODS IN BIOINFORMATICS	13
4.1	General considerations	13
4.1.1	Introduction and objectives	13
4.1.2	Problem formulation	14
4.1.3	Data mining	15
4.2	Descriptive data mining	16
4.3	Predictive data mining	17
4.4	Data mining algorithms	18
4.4.1	Clustering	18
4.4.1.1	K-means	20
4.4.1.2	SOM	20
4.4.1.3	Hierarchical clustering	21
4.4.2	Graphic based methods	22
4.4.3	Pattern Discovery	23
4.4.3.1	Apriori	23
4.4.3.2	FP-growth	24
4.4.3.3	LPMiner	24
4.4.4	K-Nearest Neighbors	25
4.4.5	Decision Trees	25
4.4.6	Artificial Neural Networks	27
4.4.7	Support Vector Machines	28
4.4.8	Bayesian Networks	28
4.4.9	Fuzzy Logic	30
4.5	Model Validation	30
4.6	Model Analysis	31

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	3/61

4.7	Meta-Modeling.....	31
4.8	Biomarkers Fusion	31
5	BIOINFORMATICS DATA INFRASTRUCTURE - DATA MANAGEMENT AND INTEGRATION	33
5.1	Introduction.....	33
5.2	Concept	33
5.2.1	The Phenotype Database (dbNP)	33
5.2.2	Data Infrastructure for Applying Models ON Design and Safety (Diamonds)	34
6	IMPLEMENTATION OF BIG DATA ANALYTICS – GENESPRING	37
6.1	General characteristics	37
6.2	Multi-omic Pathway Analysis	39
6.3	NLP Network Analysis	40
6.4	Meta-data framework.....	41
6.5	Transcriptomic analysis	42
6.6	Genomic copy number analysis	43
6.7	Genome-wide association analysis	44
6.8	Statistical tools for testing differential expression.....	44
6.9	Extensible functionality with Jython and R	45
6.10	Report Generation Capability.....	45
6.11	Modules and file formats	46
7	APPLICATION OF BIG DATA ANALYTICS - EXPOSURE TO REAL LIFE AMBIENT AIR MIXTURE	48
8	REFERENCES	56

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	4/61

Document Information

Grant Agreement Number	ENV-603946	Acronym	HEALS
Full title	Health and Environment-wide Associations based on Large population Surveys		
Project URL	http://www.heals-eu.eu/		
EU Project Officer	Tuomo Karjalainen,- Tuomo.KARJALAINEN@ec.europa.eu		


Deliverable	Number	7.1	Title	Data infrastructure and data mining model of internal exposome
Work Package	Number	7	Title	Novel bioinformatics for predictive biomarker discovery

Delivery date	Contractual	M18	Actual	28/04/2015
Status	Draft <input type="checkbox"/>		Final X	
Nature	Demonstrator <input type="checkbox"/>	Report X	Prototype <input type="checkbox"/>	Other <input type="checkbox"/>
Dissemination level	Confidential <input type="checkbox"/>		Public X	

Author (Partners)	Denis A. Sarigiannis, P. Kontoroupi, S. Karakitsios (AUTH) R. Stierum, E. van Someren (TNO)			
Responsible Author	Denis A. Sarigiannis		Email	denis@eng.auth.gr
	Partner	AUTH	Phone	+30-2310-994562

Document History

Name	Date	Version	Description

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	5/61

1 Summary

A key novelty in HEALS is the integrated use of advanced computational tools supporting environmental and biological data analyses for comprehensive data interpretation. These tools include Physiology-Based BioKinetic models (PBBK), novel bioinformatics strategies for biomarker prediction and advanced multivariate statistics for associating the links between exposure to environmental stressors and health status and investigating causality.


The current report presents the methodological framework to be followed for data mining model and computational infrastructure for use in the internal exposome, i.e. in the analysis of biomarkers (chemical, biochemical and genetic) data and storage/handling/sharing of this information seamlessly with the other parts of the HEALS database platform. In practice, the current document presents the workflow for integrating the omics information produced from human biosamples using the proper algorithms for biomarkers identification and prediction, as well as for setting the basis for the causal associations between environmental exposure and disease.

The introduction section provides the overview of the HEALS concept and the role that omics technologies. The key points are i) the way that from an untargeted analysis we move to a targeted one and ii) the way transcriptomics and metabolomics are coupled for joint pathway analysis.

The next chapter presents the different algorithms that are used in the bioinformatics science for the interpretation of the omics results, focusing on advanced data mining analysis techniques aiming at biomarkers discovery. Currently available bioinformatics will be applied and enhanced to select the most relevant omics data for given exposure/disease pathways. For descriptive data mining, the FPGrowth and LPMIner algorithms will be used for pattern discovery, whereas for data clustering a number of available tools (K-means, self-organizing maps (SOM), graph-based clustering) will be used. For predictive data mining techniques such as artificial neural networks (ANN), decision trees, support vector machines (SVM), K-nearest neighbors and Bayesian networks will be tested and integrated to improve upon the current state of the art. The DIAMONDS data infrastructure, an integrated metadata/data and analysis infrastructure for computational chemistry and genomics will be used as bioinformatics data integrator.

In addition, the implementation of big data analytics workflow guidance is given.

Finally, an example of exposure to a real life mixture, the relevant omics responses and how this is translated into pathway analysis is given, providing an overview of the approach that will be applied in HEALS in terms of associating molecular biology techniques to risk assessment and population health effects.


 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	6/61

2 Introduction - General considerations

The exposome (Wild, 2005) represents the totality of exposures from conception onwards, simultaneously identifying, characterizing and quantifying the exogenous and endogenous exposures and modifiable risk factors that predispose to and predict diseases throughout a person's life span. Exposome came as a complement to the human genome; although decoding of human genome (Schmutz et al., 2004) increased our understanding of the underlying causes of disease, genome explains only a percentage of population burden. Thus, it is evident that environmental factors are equally or eventually more important and what is actually critical is the interaction of environmental factors with the biological systems. Towards a better understanding of the causal links among genome, environment and disease, unraveling the exposome implies that both environmental exposures and genetic variation are reliably measured simultaneously.

HEALS (Health and Environment-wide Associations based on Large population Surveys) brings together a comprehensive array of novel technologies, data analysis and modeling tools that support the efficient design and execution of large-scale exposome studies. The HEALS approach brings together and organizes environmental, socio-economic, exposure, biomarker and health effect data; in addition, it includes all the procedures and computational sequences necessary for applying advanced bioinformatics coupling advanced data mining, biological and exposure modeling so as to ensure that environmental exposure-health associations are studied comprehensively. The overall approach will be verified in a series of population studies across Europe, tackling various levels of environmental exposure, age windows and gender differentiation of exposure, and socio-economic and genetic variability. The main objective of HEALS is the refinement of an integrated methodology and the application of the corresponding analytical and computational tools for performing environment-wide association studies in support of EU-wide environment and health assessments. For the first time, HEALS will try to reverse the paradigm of "nature versus nurture" and adopt one defined by complex and dynamic interactions between DNA sequence, epigenetic DNA modifications, gene expression and environmental factors that all combine to influence disease phenotypes. HEALS will start from analysis of data collected in on-going epidemiological EU studies involving mother/infant pairs, children, or adults including the elderly to evidence relevant environmental exposure/health outcome associations. These associations will aid in designing pilot surveys using an integrated approach, where the selection of biomarkers of exposure, effects and individual susceptibility results in integrated risk assessment.

The overall methodological concept of HEALS and the different arrays involved is graphically illustrated in Figure 1. This includes a wide array of state of the art technologies across all major disciplines of the environmental exposure, biochemistry, molecular biology, toxicology, bioinformatics and epidemiology arena.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	7/61

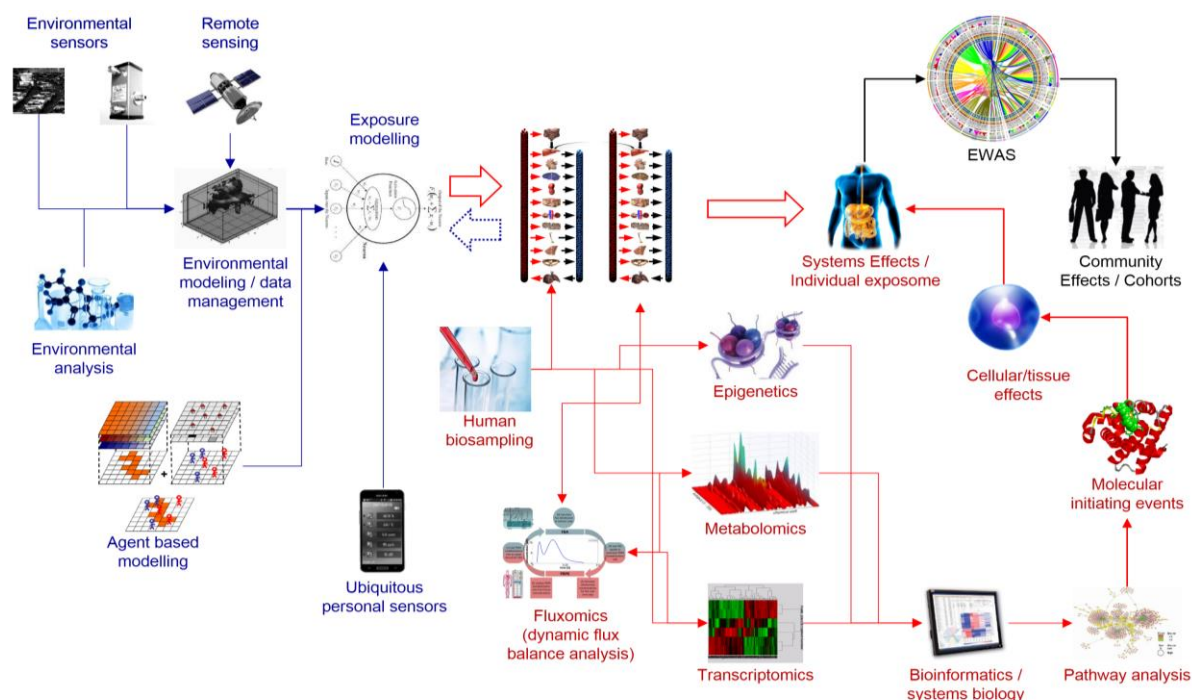



Figure 1: Overall HEALS conceptual methodological framework

Another key aspect of HEALS is the development of innovative **bioinformatics** strategies for biomarker prediction. The bioinformatics tools currently available for biomarker detection and analysis range from statistical approaches to data mining. The latter is the process of discovering valuable information from large amounts of data in the form of associations, patterns, changes, or significant structures. Data mining can be *descriptive* or *predictive*. The distinction between descriptive and predictive data mining models is in many cases unclear mainly because the same tools can be used either way. However, when prediction is under consideration, it can be reached through *classification* or *regression*. The available techniques for both tasks are numerous, with decision trees, neural networks and support vectors more widely used. An additional issue to be taken under consideration related to exposome studies is “big data”. This is the result of the huge amount of information at various levels of biological organization reflected as genome, transcriptome, proteome, metabolome accompanied by other population relevant information (exposure, medical status). The name itself suggests huge amount of data, which, however, represents only one aspect. In general, big data has four important features, so called four V’s (Li and Chen, 2014): volume of data, velocity of processing the data, variability of data sources, and veracity of the data quality.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	8/61

3 Methodological framework for big data analytics

3.1 Omics workflow

The HEALS approach to the internal exposome relies on evaluating the maximum available information from multiple omics data. Towards this aim, a well-structured exposure biology workflow has to be followed, as graphically illustrated in Figure 2.

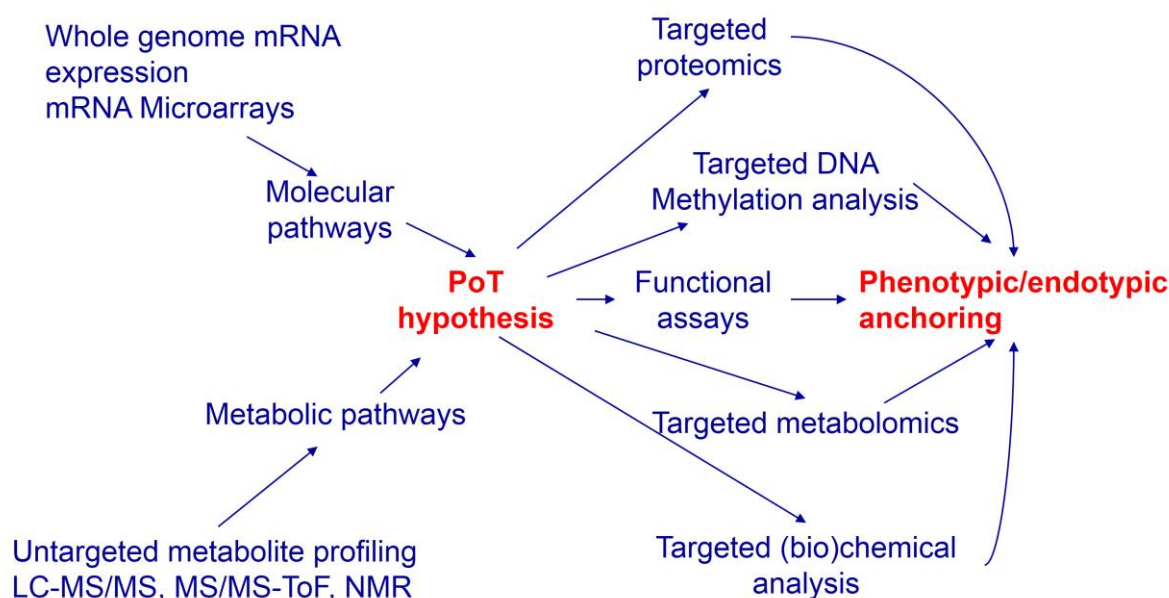



Figure 2: From untargeted to targeted omics workflow

This workflow is specifically designed to decipher the causal links of genome, environment and disease, entailing a stepwise process:

1. The overall assessment needs to be initiated from an agnostic point of view; no former hypothesis has to be formed unless driven by some evidence of perturbed pathways. Thus, the first step of the assessment is the investigation of untargeted –omics, including whole genome mRNA expression and untargeted metabolite profiling.
2. The results of untargeted metabolomics and transcriptomics will be jointly analysed (molecular and metabolic pathway analysis) so as to construct putative hypotheses on Pathways of Toxicity (or of Adverse Health Outcome).

After building up the putative hypothesis of the Pathways of Toxicity (thus limiting the number of potential pathways to be perturbed), additional evidence will be provided through targeted –omics (including proteomics and metabolomics), DNA methylation, functional assays and biochemical biomonitoring. Anchoring of the different endotypes will be further supported by *in vitro* assays for mechanistic confirmation, biology-inspired modeling and bioinformatics for data analysis.


This will result in:

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	9/61

- (a) Identification of endotypes (early biological event markers) that can be associated causally with phenotypes of adverse health outcomes.
- (b) Determination of allostatic status of the analysed individuals
- (c) Association of allostatic status with individual risk of disease and eventually burden of disease on the population level

Technically, the different aspects of analytical exposure biology applied in HEALS include: (a) transcriptomics, meaning the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods, such as microarray analysis (b) Metabolomics (untargeted and targeted); entails the global analysis of metabolites in easily accessible human body fluids. Established and robust sample preparation protocols (e.g. for plasma, urine) in combination with the latest generation of NMR and ultra-high-pressure liquid chromatography (UPLC)-mass spectrometers (MS), high-pressure liquid chromatography-electrochemical detection (HPLC-ED), and LC-ECA (electrochemical coulometric arrays) will be applied. For novel matrices (e.g. meconium) ad hoc protocols have been developed. These approaches will be applied to samples from selected human cohorts to determine the impact of exposure to chemicals (e.g. pesticides, industrial solvents etc) on the human exposome. (c) Adductomics; deals with the measurement of adducts of electrophiles with DNA, blood proteins (haemoglobin, albumin), and glutathione. HPLC-ED and MS will be used to evaluate levels of 8-hydroxyguanine and O6-alkylguanine in blood samples from human populations, in relation to genetic susceptibility (DNA repair genotype). MS based technology (GS-MS) will be employed to quantify blood protein adducts. Also, LC-MS/MS and fixed-step selected reaction monitoring technologies (FS-SRM) for measuring protein adducts, will be further developed for application in HBM settings to contribute to the definition of the exposome. With respect to SNP profiling in toxicologically relevant genes, genotyping assays will be developed and optimized using Taqman technology on the existing Biomark Fluidigm platform. In addition, functional assays of specific DNA repair proteins will be performed. Genome-wide DNA methylation profiling and miRNA expression are key to investigating the importance of epigenetic effects on environmental health. Array- and sequencing-based DNA methylation profiling technologies will be used to meet this goal. In particular, Agilent microarray miRNA profiling and genome-wide bisulfite sequencing will be used to analyse miRNA expression and DNA methylation respectively, on the biosamples of the cohorts:

- i) Apply SNP analysis, miRNA analysis and next-generation sequencing to define genetic susceptibility for chemicals at population level (e.g. DNA repair phenotypes, Phase II reaction genotypes)
- ii) Identify differences between epigenetically influenced and independent SNPs
- iii) Develop sample handling and metabolomics workflows
- iv) Generate metabolite expression data; identify biomarkers from cohorts and *in vitro* models (glutathione, s-adenosyl methionine, bisphenol A)
- v) Develop methodologies (FS-SRM) and generate DNA and protein adduct data from cohorts.
- vi) Determine methyl/hydroxymethylcytosine at specific genomic sequences (promoters, CpG islands, repeated sequences)


 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	10/61

- vii) Identify Illumina methylation at 20 genes, and apply Methylator phenotypic analysis for this set of genes.

The causal link between exposure and disease endpoint as revealed by the identification of –omics markers of effect requires mechanistic confirmation, which will be based on *in vitro* assays assessing complex toxic endpoints relevant to disease endpoints on cell models derived from liver (hepG2/HuH7, HepaRG), adipocytes (hMADS) or cells differentiating into neurons (C11) to establish for the different technologies omics compound specific hazard signatures. This will also support the system biology model of causation developed in the bioinformatics part of the HEALS. Next to confirming the systems biology exposome model, these mechanistic studies provide kinetic constants and compound receptor-binding data useful to further develop the model. In this context, it is possible to use these data to design *in vitro* human cell-based assays and systems biology experiments, to mechanistically anchor omics observations from cohort studies in light of the linkage between exposure and health outcome data. The hypothesis is put forward that the correspondence of omics signatures in cohort biomaterials with those obtained from experimental *in vitro* models can demonstrate the validity of chosen biomarkers. In fact this approach can be used to corroborate if: (a) signatures are more closely related to the actual compound exposure (e.g. biotransformation) and thus represent a marker for exposure; (b) more closely related to mechanisms of disease, and consequently represent a marker for health impact. Overall, empirical perturbation detected from HBM data (including omics markers) can be integrated into *in vitro* testing to help verify system biology-level or “real-world” human toxicity of environmental stressors (Pleil, 2012).

3.2 Coupling *in vitro* and *in silico* analysis

A major concept that has to be highlighted herein is the joint analysis of transcriptomics and metabolomics. We need to keep in mind that none of these type of omics technologies could be a standalone solution for deploying exposome, meaning to deploy a causative association between exposure and disease. Upon exposure to an external stressor, it is expected that many genes will be differentially expressed (upregulated or downregulated). However, although many metabolic pathways might be potentially altered, it is well known that a small number of differentially expressed genes will result in potentially perturbed pathways. On the other hand, metabolomics represents the outcome of an already modified metabolic process. However, metabolomics alone does not provide information on which was the perturbed pathway that resulted in observable change in the metabolic profiles. Joint pathway analysis coupling transcriptomics and metabolomics data (Figure 3), allows the identification of the missing link for connecting the responses at different levels of biological organization and thus, linking environmental exposure to disease.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	11/61

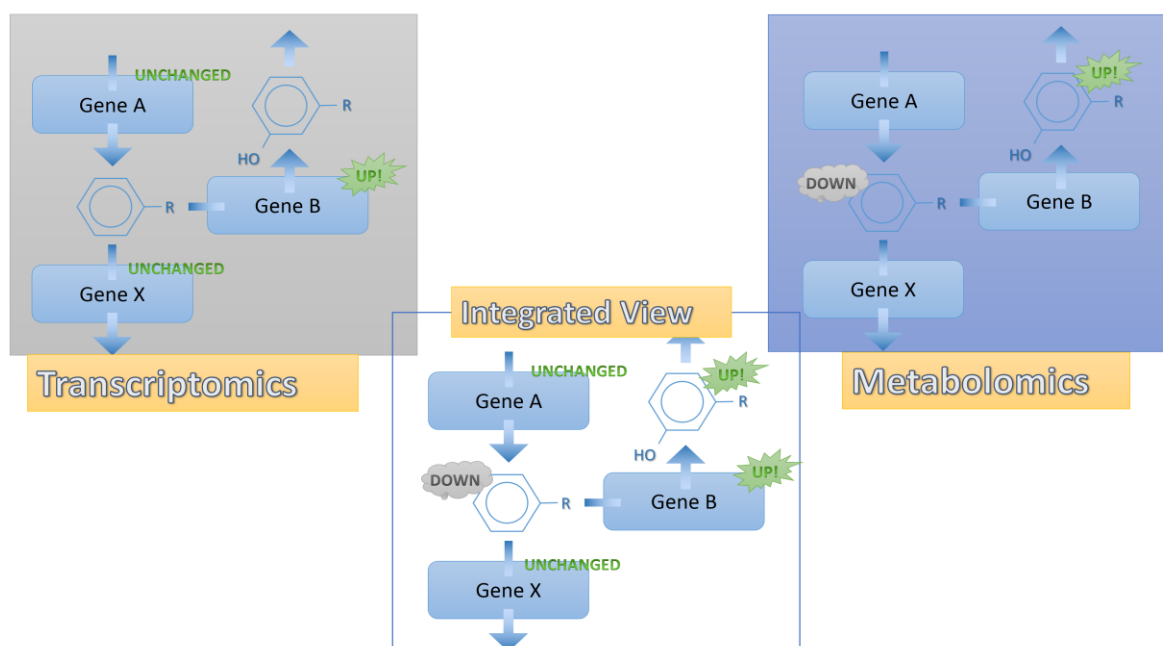


Figure 3: Integrated view of transcriptomics and metabolomics responses

The array of technologies to support the approach are graphically illustrated in Figure 4. These include both in vitro analytical techniques, as well as in silico approaches for big data analytics.

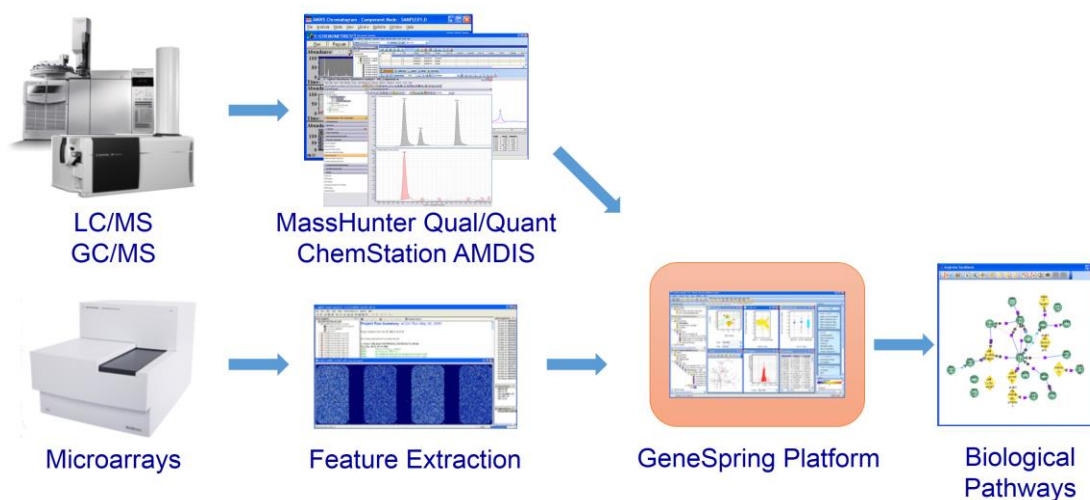



Figure 4: LC/MS and GC/MS for metabolomics analysis, microarrays for gene expression, and dedicated software (Genespring) for joint pathway analysis

Interpretation of multi-omics data will require information from several relevant multi-omics and pathway analysis databases; an overview of these databases and the respective information workflow is given in Figure 5.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	12/61

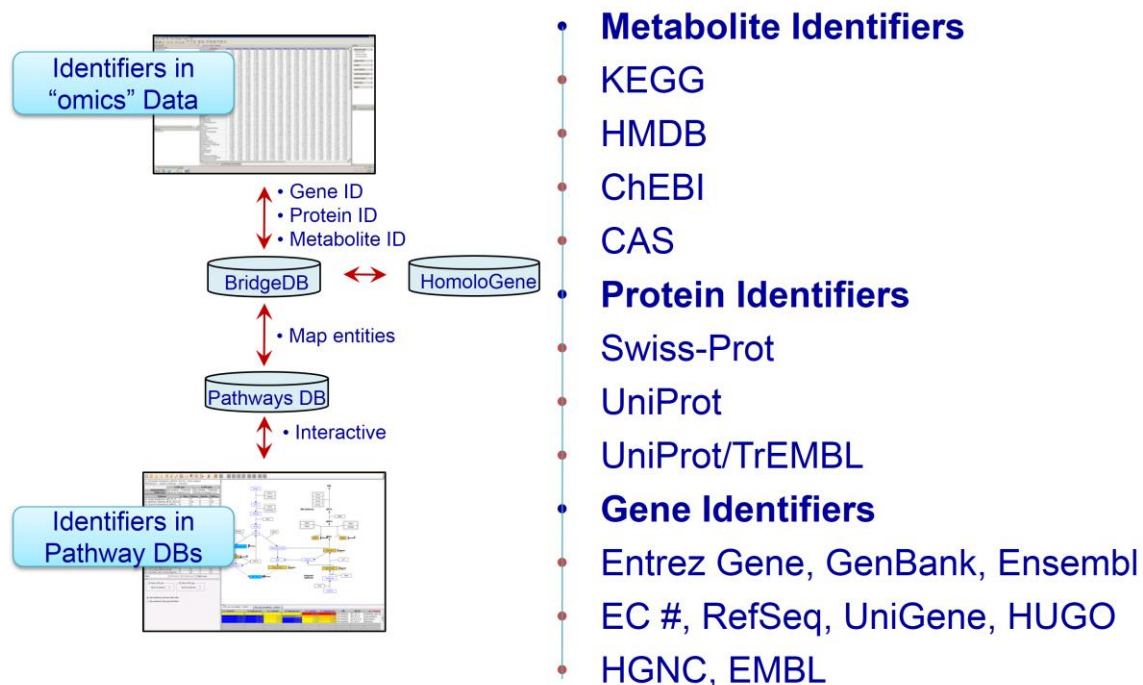



Figure 5: Information flow among omics and pathways databases

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	13/61


4 Statistical methods in bioinformatics

4.1 General considerations

4.1.1 Introduction and objectives

The scope of the available statistical methods is to understand the biological functions of toxicity pathway interactions in relation to external/internal exposure, to confirm the causative effect between exposure and disease endpoint through theoretical (computational) models, to combine mixed data, resulted from various sources, through the utilization of advanced data mining analysis techniques, to provide the methodological tools for integrating multiple biomarkers into a mechanistic description and to derive the systems biology exposome model.

Gene expression data, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks, and cellular states. One goal in analyzing expression data is to try to determine how the expression of any particular gene might affect the expression of other genes; the genes involved in this case could belong to the same gene network. By a gene network, we mean a set of genes being expressed together in a non-random pattern. Another goal of expression data analysis is to try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells. While early experiments using microarrays profiled only a few samples, more recent experiments profile on the order of dozens or even hundreds of samples, allowing for a more robust statistical analysis of the data. In the near future, data sets containing thousands of samples should become available. As gene expression data sets become larger and larger, spreadsheets will become less and less of an adequate tool for doing and data mining techniques using large databases should find more and more use in analyzing expression data. In the analysis of gene expression data, the items in an association rule can represent genes that are strongly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. a diagnosis for a tumor sample that was profiled, or a drug treatment given to cells in the sample before profiling). An example of an association rule mined from expression data might be [cancer] => gene A↑, gene B↓, gene C↑, meaning that, for the data set that was mined, in most profile experiments where the cells used were cancerous, gene A was measured as being up (i.e. highly expressed), gene B was down (i.e. highly repressed), and gene C was up, altogether. Hence scope is to interpret gene expression technology results via integration of gene expression profiles with corresponding biological knowledge (gene annotations, literature, etc.) extracted from biological databases. Consequently, the key task in the interpretation step is to detect the present co-expressed (sharing similar expression profiles) and co-annotated (sharing the same properties such as function, regulatory mechanism, etc.) gene group. Several approaches dealing with the interpretation problem have recently been reported. These approaches can be classified in three axes (Chang, 2002; Martinez, 2007): expression-based approaches, knowledge-based approaches and co-clustering approaches. The most currently used interpretation axis is the expression-based axis that gives more weight to gene expression profiles. However, it presents many well-known drawbacks. First, these approaches cluster genes by similarity in expression profiles across all biological conditions. However, gene groups involved in a biological process might be only co-expressed in a small subset of conditions (Altman and Raychaudhuri, 2001). Second, many genes have different biological roles in the cell, they may be conditionally co-expressed with different groups of genes. Since almost all clustering methods used place each gene in a single cluster, that is a single group of genes, relationships with different groups of conditionally regulated genes may remain undiscovered (Gasch, 2002). Third,

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	14/61

discovering biological relationships among co-expressed genes is not a trivial task and requires a lot of additional work, even when similar gene expression profiles are related to similar biological roles (Shatkay, 2000).


With regard to the objectives of HEALS, bioinformatic techniques will be applied and enhanced in order to select the most relevant omics data and derive specific data profiles for given exposure/disease pathways. In particular for the WP7, predictive-biomarkers will be determined, based on heterogeneous datasets, resulting from human biomonitoring, omics and epigenetics analyses and PBBK modeling. This would require, firstly the pre-processing of the data produced, secondly the discovery of specific data patterns and/or clusters, thirdly the creation of data models based on training sets and, finally, the evaluation of the models with regard to their validity and prediction capacity on the basis of the test data. Therefore several approaches will be implemented from the fields of descriptive and predictive data mining to achieve our goals. Their results will be systematically assessed and the model that best describes our exposome data will be employed to the subsequent population surveys. Multiple omics biomarkers will also be integrated to a mechanistic description of toxicity pathway interactions, in relation to external/internal exposure, achieved by developing systems biology pathway models and using the predictive bioinformatics approaches. In this regard statistical algorithms will be implemented for HEALS in task 7.1 (i.e. the descriptive data mining – preprocessing, data clustering and pattern discovery), task 7.2 (i.e. the predictive data mining – data models design and analysis), task 7.3 (i.e. the model integration – biomarkers identification and prediction validation) and task 7.4 (i.e. the bioinformatics data infrastructure for storage of human cohort study specific metadata in relation to omics and (bio)assay data).

4.1.2 Problem formulation

Before applying any unsupervised learning functions to the HEALS datasets and identifying intrinsic relations on them, it will be necessary to adopt first a pre-processing step, which will consist of five essential modules:


- technology specific data pre-processing (e.g. spectra de-convolution)
- noise removal, to ensure the consistency and high quality of our data from possible outliers or discrepancies in the measurements
- data transformation, to normalize the values in our dataset and also increase their generalization,
- data reduction, to decrease both the apparent complexity in our data, through subset representations, and the dimensionality of the derived models, and
- discretization, to scale the data and prepare them for further analysis by means of clustering and pattern extraction.

After the pre-processing step, during the modeling procedure, training and testing of the various datasets will follow the learning process. This is required in order to depict relations among the observed variables, to recognize complex patterns and to extract rules. Lastly, since the number of states required to be recognized may be much larger than the available dataset; a representative training data set will be selected with aim to produce a possible generalization to the entire data set.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	15/61

4.1.3 Data mining

Data mining (Fayyad, 1996) systems combine techniques from many research areas, e.g., statistics or computer science (database systems and machine learning). The systems can be categorized according to the task they solve. There are, e.g., classification, regression, clustering, or descriptive systems. In addition the systems can be also characterized by the type of knowledge they produce. In this regard there are connectionistic (artificial neural networks), statistics (Naïve Bayes classifier), or logic (decision trees or classification rules) systems. Some algorithms combine several methods such as Domingos' RISE algorithm (Dong, 1999) which integrates instance-based and rule-based learning. Lastly, the data complexity criteria separate data mining systems into two groups: propositional and relational systems. Although statistics or connectionistic systems have achieved good results in text mining, i.e., text classification and categorization, they may not be suitable for mining knowledge intended for further analysis. Nevertheless, it may be very difficult to process a complex data, such as data describing chemical molecules (structure of organic molecules) and biological (strings of DNA) data. Following the axiom that "*a data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of models or patterns*", the data mining algorithms can be classified in accordance to Figure 6. Accordingly, model structures are categorized between the prediction types (incl. linear regression, piecewise linear, nonparametric regression and classification), the probability distributions (incl. parametric models, mixtures of parametric models, graphical markov models) and the structured types (incl. time series, markov models, mixture transition distribution models, hidden markov models). In this report, the data mining algorithms will be categorized between two types: the descriptive and the predictive. The former describe the dataset in a concise and brief manner and present general data properties. The latter performs inference on the available dataset and predicts the outcome of the new datasets via the generation of one or more models.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	16/61

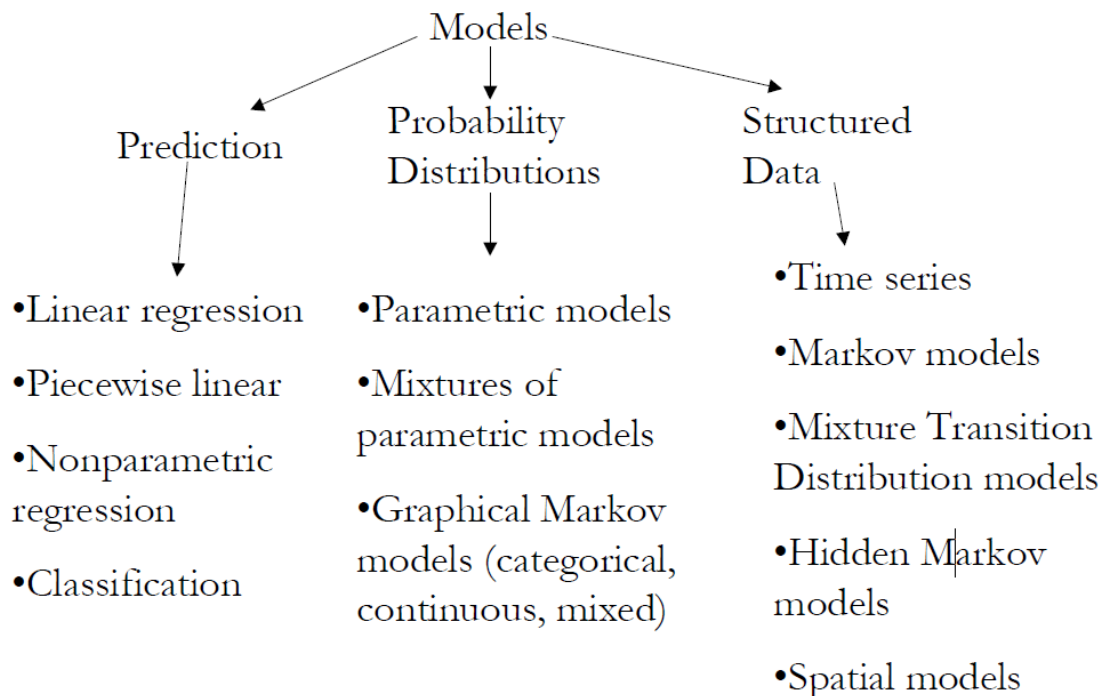



Figure 6: Model classification scheme

4.2 Descriptive data mining

A descriptive data mining approach presents the main features, where data are randomly generated from a “good” descriptive model that has the same characteristics as the ‘real’ data. In the descriptive data mining approach, patterns are evaluated either globally or locally, in accordance to the classification presented in Figure 7. Global patterns include clustering methods via portioning, hierarchical clustering and mixture modelling and local patterns including the outlier detection, the changepoint detection pattern, the ‘bump’ hunting, the scan statistics and the association rules.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	17/61

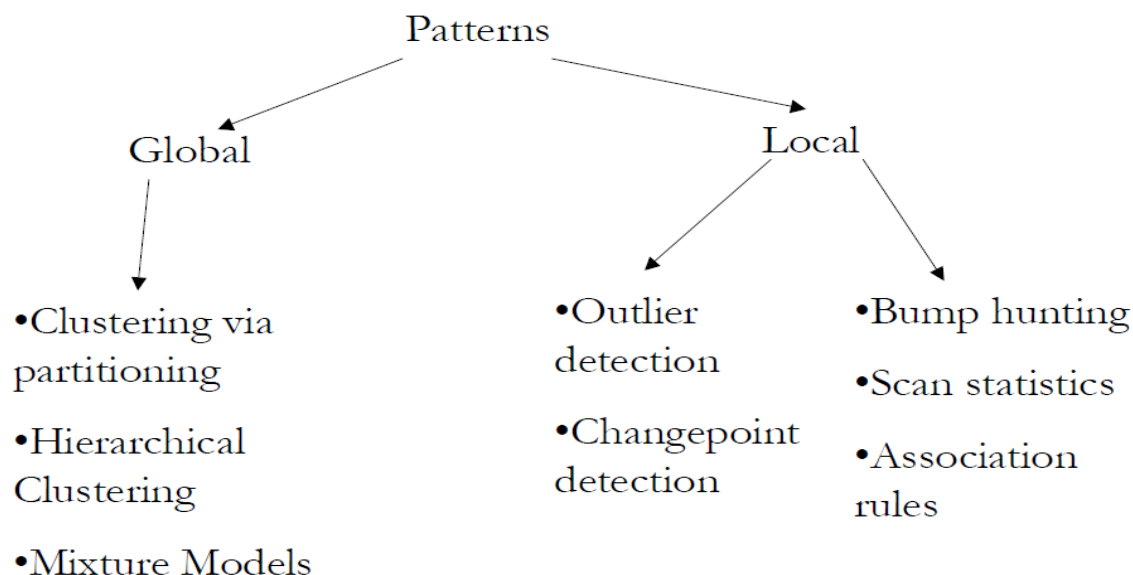



Figure 7: Classifications in the descriptive data mining approach

4.3 Predictive data mining

Predictive data mining can be used to forecast explicit values, based on patterns determined from known results. Several techniques can be used for that purpose, ranging from typical approaches based on decision trees or k-nearest neighbors to more sophisticated ones that employ Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) or Bayesian Networks (BNs). Hence it is possible to perform inference on the available dataset, to perform a mere classification and to study and unravel the feature attributes concealed in data (e.g. exposome). Various computational techniques, especially machine learning algorithms (Larranaga, 2003), are applied, for example, to select genes or proteins associated with the trait of interest and to classify different types of samples in gene expression of microarrays data (Allison et al., 2006) or Mass Spectrometry (MS)-based proteomics data (Aebersold, 2003), to identify disease associated genes, gene-gene interactions, and gene-environmental interactions from Genome Wide Association (GWA) studies (Hirschhorn and Daly, 2005), to recognize the regulatory elements in DNA or protein sequences (Zeng et al., 2009), to identify protein-protein interactions (Valencia and Pazos, 2002), or to predict protein structure (Jones, 2001). The aim of designing/using ensemble methods (Breiman, 1996; Breiman, 2001; Freund, 1996) is to achieve more accurate classification (on training data) as well as better generalization (on unseen data). However, this is often achieved at the expense of increased model complexity (i.e. decreased model interpretability) (Kuncheva, 2004). A better generalization property of ensemble approach is often explained using the classic bias-variance decomposition analysis (Webb, 2004). Specifically, previous studies pointed out methods like bagging (Fig. 8(a)) that improve generalization by decreasing variance (Breiman, 1998) while methods similar to boosting (Fig. 8(b)) achieve this by decreasing bias (Schapire, 1998).

 FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	18/61

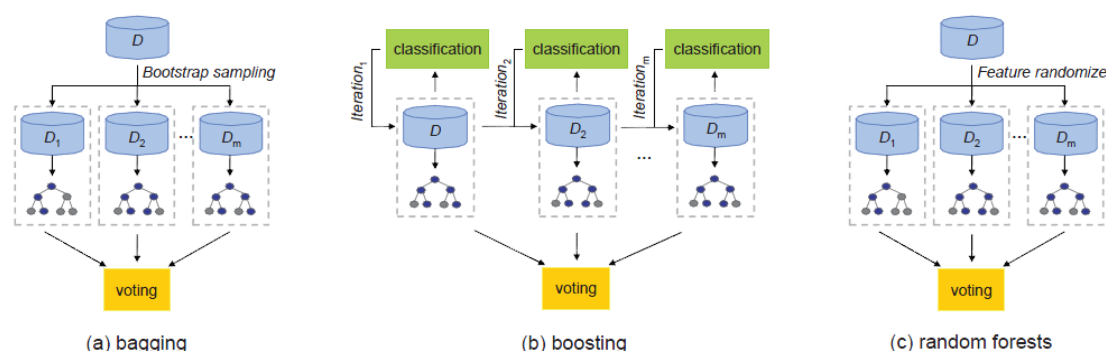



Figure 8: Schematic illustration of the three popular ensemble methods.

4.4 Data mining algorithms

4.4.1 Clustering

DNA microarray technology has made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. In this regard, clustering techniques can be utilized to reveal the natural structures and identify the patterns in the underlying data. Specifically, cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. Many conventional clustering algorithms are available to be adapted or directly applied to the gene expression data. Assuming that a 2d matrix of the gene expression is available, the row will represent the genes and the columns will represent different experiments. This representation corresponds to a gene expression profile, available for clustering. Figure 9 illustrates such an example, where rows represent the genes and columns people with leukemia. The scope is to organize profiles into clusters so that the instances in the same cluster are highly similar to each other and the instances from different clusters have low similarity to each other.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	19/61

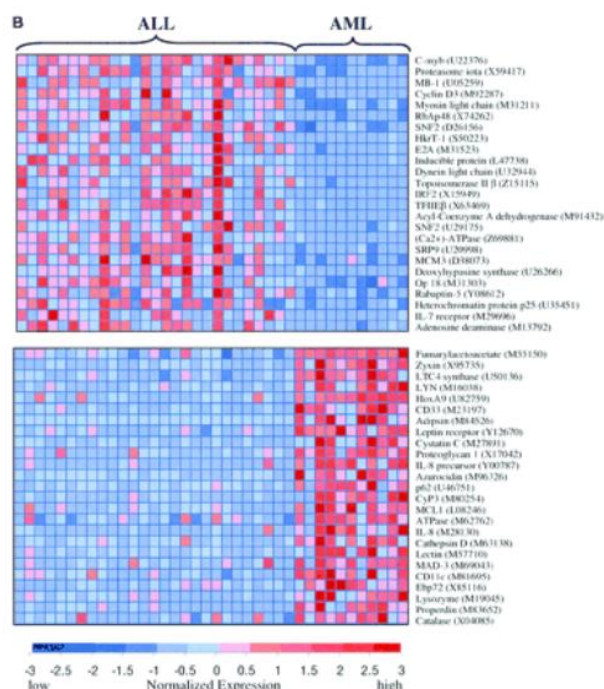



Figure 9: A sample expression profile

Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene-based clustering presents several challenges. Firstly, cluster analysis is typically in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis. For example, a clustering algorithm which can accurately estimate the “true” number of clusters in the data set would be more favored than one requiring the pre-determined number of clusters. Secondly, due to the complex procedures of microarray experiments, gene expression data often contain a huge amount of noise. Therefore, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise. Thirdly, gene expression data are often “highly connected” (Jiang et al., 2003), and clusters may be highly intersected with each other or even embedded one in another (Jiang, 2003). Therefore, algorithms for gene-based clustering should be able to effectively handle this situation. Finally, users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters (e.g., which clusters are more close to each other, and which clusters are remote from each other) and the relationship between the genes within the same cluster (e.g., which gene can be considered as the representative of the cluster and which genes are at the boundary area of the cluster).

In this regard the available methods include the K-means algorithm (McQueen, 1967), the Self-organizing Map (SOP) method (Kohonen, 1984), the Hierarchical clustering algorithms, the Graph-theoretical approaches (Ben-Dor, 1999; Shamir, 2000), the pattern clustering methods (Agrawal, 1994; Han, 2003; Seno, 2001) and the model based clustering (Dasgupta and Raftery, 1998; Fraley and Raftery, 1998).

 FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	20/61

4.4.1.1 K-means

The K-means algorithm (McQueen, 1967) is a typical partition-based clustering method. Given a pre-specified number K , the algorithm partitions the data set into K disjoint subsets which optimize the following objective function 1,

$$E = \sum_{i=1}^K \sum_{O \in C_i} |O - \mu_i|^2 \quad (1)$$

Where O is a data object cluster C_i and μ_i is the centroid (mean of objects) of C_i . Hence according to the objective function 1, E is the minimized sum of the squared distances of objects from their cluster centres. The time complexity of the K-means method is $O(l*k*n)$, where l is the number of iterations and k is the number of clusters. While the K-means cluster converges after a small number of iterations, with regard to the gene-based clustering algorithm, several drawbacks are identified: firstly, the number of gene clusters in a gene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of k and compare the clustering results. For a large gene expression data set which contains thousands of genes, this extensive parameter fine-tuning process may not be practical. Secondly, gene expression data typically contain a huge amount of noise; however, the K-means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise (De Smet et al., 2002; Sherlock, 2000). Other clustering algorithms available to overcome the drawbacks of the K-means algorithms include those that typically use some global parameters to control the quality of resulting clusters (e.g., the maximal radius of a cluster and/or the minimal distance between clusters). However, the qualities of clusters in gene expression data sets may vary widely. Thus, it is often a difficult problem to choose the appropriate globally-constraining parameters. The implementation of the K-means is depicted in Figure 10.

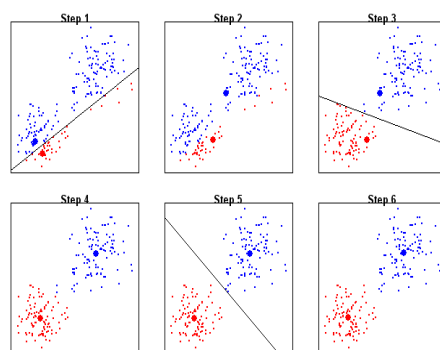



Figure 10: Implementation of the K-means algorithm to a data set

4.4.1.2 SOM

The Self-Organizing Map (SOM) method (Kohonen, 1984) was developed, on the basis of a single layered neural network. Accordingly, the data objects are presented at the input, and the output neurons are organized with a simple neighborhood structure such as a two dimensional $p * q$ grid. Each neuron of the neural network is associated with a reference vector and each data point is “mapped” to the neuron with the “closest” reference vector. In the process of running the algorithm,

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	21/61

each data object acts as a training sample which directs the movement of the reference vectors towards the denser areas of the input vector space, so that those reference vectors are trained to fit the distributions of the input data set. When the training is complete, clusters are identified by mapping all data points to the output neurons. The SOM method generates an intuitively-appealing map of a high-dimensional data set in 2D or 3D space and places similar clusters near each other. The neuron training process of SOM provides a relatively more robust approach than K-means to the clustering of highly noisy data (Herrero, 2001; Tamayo, 1999). However, SOM requires users to input the number of clusters and the grid structure of the neuron map. These two parameters are preserved through the training process; hence, improperly-specified parameters will prevent the recovering of the natural cluster structure. Furthermore, if the data set is abundant with irrelevant data points, such as genes with invariant patterns, SOM will produce an output in which this type of data will populate the vast majority of clusters (Herrero, 2001). In this case, SOM is not effective because most of the interesting patterns may be merged into only one or two clusters and cannot be identified. Figure 11 illustrates how from a 4x4 array of neurons and a random weight initialization, the weights values aid to evaluate the final data profile grouping.

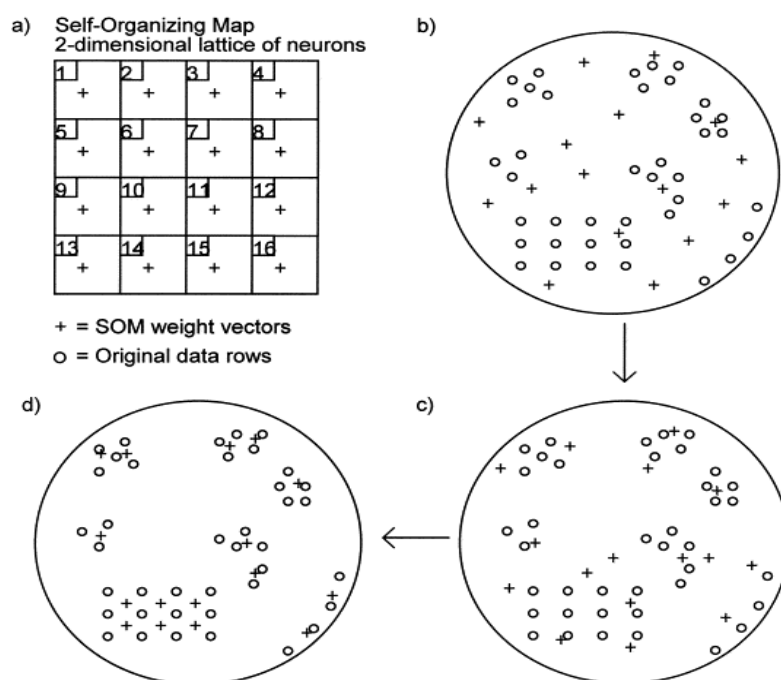



Figure 11: Implementation of SOM to data profiles

4.4.1.3 Hierarchical clustering

Hierarchical clustering generates a hierarchical series of nested clusters which can be graphically represented by a tree, called *dendrogram*, as seen in **Errore. L'origine riferimento non è stata trovata..** The branches of a *dendrogram* not only record the formation of the clusters but also indicate the similarity between the clusters. By cutting the *dendrogram* at some level, we can obtain a specified number of clusters. By reordering the objects such that the branches of the corresponding *dendrogram* do not cross, the data set can be arranged with similar objects placed together. Hierarchical clustering algorithms can be further divided into agglomerative approaches

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	22/61

and divisive approaches based on how the hierarchical *dendrogram* is formed. Agglomerative algorithms (bottom-up approach) initially regard each data object as an individual cluster, and at each step, merge the closest pair of clusters until all the groups are merged into one cluster. Divisive algorithms (top-down approach) starts with one cluster containing all the data objects, and at each step split a cluster until only singleton clusters of individual objects remain. For agglomerative approaches, different measures of cluster proximity, such as single link, complete link and minimum-variance (Dubes, 1988; Kaufman, 1990), derive various merge strategies. For divisive approaches, the essential problem is to decide how to split clusters at each step.

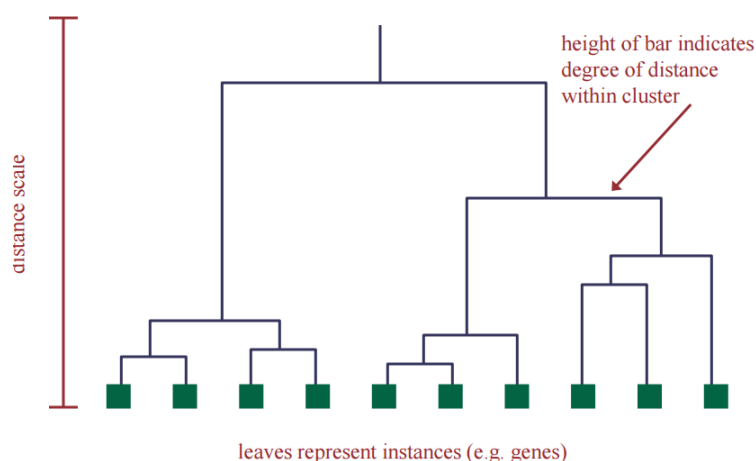



Figure 12: An example to the Hierarchical clustering

4.4.2 Graphic based methods

Using the graph-theoretical approaches, data are represented graphically using a proximity matrix P , where $P[i,j] = \text{proximity}[O_i, O_j]$ and a weighted graph $G(V,E)$ denoted as the proximity graph, where each data point corresponds to a vertex. Two general approaches are available, firstly each pair of objects is connected by an edge with weight assigned according to the proximity value between the objects (Shamir, 2000; Xing, 2001) and secondly, proximity is mapped only to either 0 or 1 on the basis of some threshold, and edges only exist between objects i and j , where $P[i,j]$ equals 1 (Ben-Dor, 1999; Hartuv, 2000). It is noted that the Graph-theoretical clustering techniques are explicitly presented in terms of a graph, thus converting the problem of clustering a dataset into such graph theoretical problems as finding minimum cut or maximal cliques in the proximity graph G . Figure 13, depicts the graph based methods. Graph vertices are grouped into clusters. From these representations the vertex similarity measures are estimated including distance, adjacency and the degree of connectivity. Other cluster fitness measures include the density, the conductance, the modularity and the centrality.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	23/61

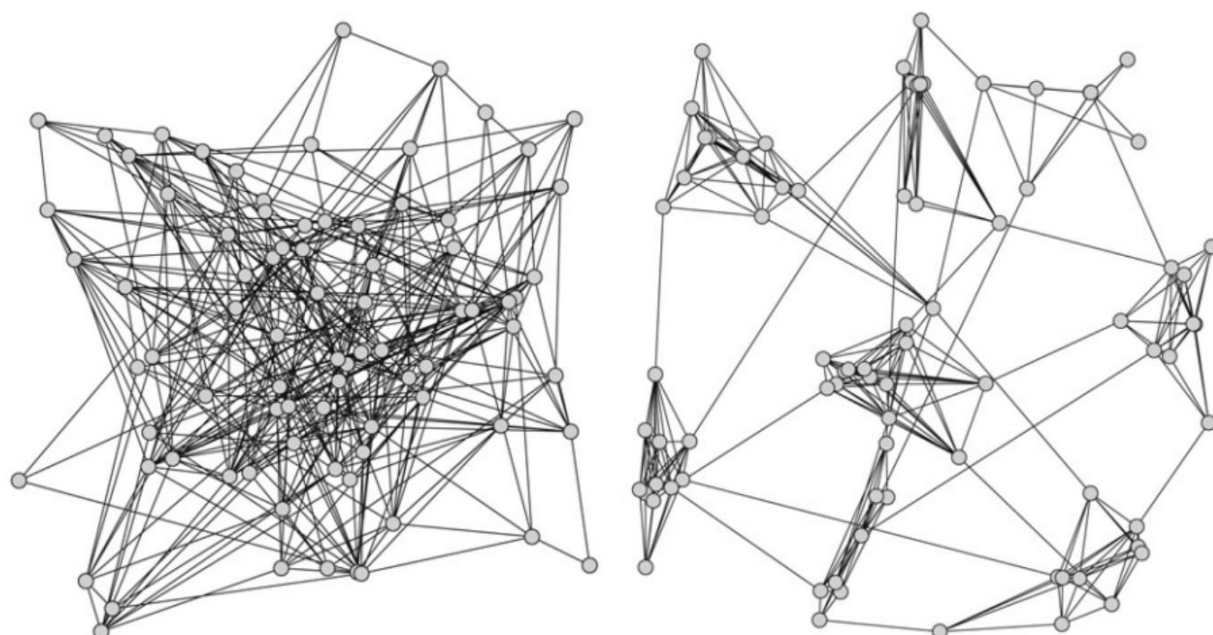


Figure 13: Global vs Local clustering in graph based methods

4.4.3 Pattern Discovery


Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. An association rule has the form $LHS \rightarrow RHS$, where LHS and RHS are disjoint sets of items, the RHS set being likely to occur whenever the LHS set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes. It is hence a commonly used methodology for detecting local patterns in unsupervised learning systems and represent feature-value conditions among the data. Analysis of large genomic data has two important goals:

- to determine how the expression of any particular gene might affect the expression of other genes; the genes involved in this case could belong to the same gene network. By a gene network, we mean a set of genes being expressed together in a non-random pattern.
- to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells.

A number of different algorithms for pattern extraction are available including the Apriori, the FP-growth and the LPMIner used either independently or in combination, described in sections 4.4.3.1 – 4.4.3.3.

4.4.3.1 Apriori

The most basic join-based algorithm is the *Apriori* method (Agrawal, 1994). The *Apriori* approach uses a level-wise approach in which all frequent item-sets of length k are generated before those of length $(k + 1)$. The main observation which is used for the *Apriori* algorithm is that every subset of a

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	24/61

frequent pattern is also frequent. Therefore, candidates for frequent patterns of length $(k+1)$ can be generated from known frequent patterns of length k with the use of *joins*. It is noted that *join* is defined by pairs of frequent k -patterns that have at least $(k - 1)$ items in common. This method identifies the frequent items in the dataset: i.e. those items with minimum *Support*. Here Support (S) is defined as the proportion of records in the dataset which contain the item set. An advantage of this approach lies on the fact that it is bottom-up approach and hence can be extended to larger item sets that appear often in the dataset. Figure 14 depicted the implementation of the Apriori method from a database D .

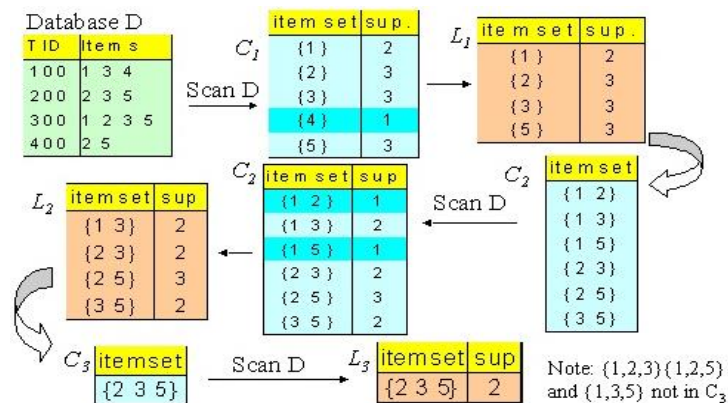



Figure 14: Implementation of the Apriori method from a database D

4.4.3.2 FP-growth

The most popular frequent itemset mining is the FP-Growth algorithm (Han, 2000). The main aim of this algorithm was to remove the bottlenecks of the Apriori-Algorithm in generating and testing candidate set. The problem of Apriori algorithm was dealt with, by introducing a novel, compact data structure, called frequent pattern tree, or FP-tree then based on this structure an FP-tree-based pattern fragment growth method was developed. FP-growth uses a combination of the vertical and horizontal database layout to store the database in main memory. Instead of storing the cover for every item in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item. This new data structure is denoted by FP-tree (Frequent-Pattern tree) (Han, 2003). Essentially, all transactions are stored in a tree data structure. Hence the advantages of this method are the that it avoids the costly iterative scans and generations of large number of candidate pattern sets and produces a more compact pattern set.

4.4.3.3 LPMiner

The LPMiner approach (Seno, 2001) is based on FP-growth method, using false frequent patterns. The main idea behind is that the minimal frequency threshold decreases with the length of the pattern then a smaller number of false frequent patterns is generated. Therefore an additional constraint was introduced, the *length-decreasing support* constraint. Using this method, frequent items are found whose support decreases as a function of the item length. This method combines the FP-tree data structure with pruning process and it is known to be faster than the FP-growth method. Nevertheless, it is problematic for the short item sets.

 FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	25/61

4.4.4 K-Nearest Neighbors

The simplest of all the machine learning algorithms, is the K-nearest algorithm, based on instance based learning. Data classification is based on the methodology of k nearest neighbors where the output is object value equal to the average of the values of their k nearest neighbors. It is noted that weights are used to regulate the attribute's contribution, when k=1 the object is assigned to the class of the nearest neighbor. Certainly the selection of the Ks is crucial and in our case will be accomplished heuristically.

4.4.5 Decision Trees

Decision trees (Breiman, 1984; Quinlan, 1986) are among the most popular learning algorithms and they have been applied extensively in computational biology. The key ingredients of the success of these methods are their interpretability that makes their model transparent and understandable to human experts, their flexibility, that makes them applicable to a wide range of problems, and their ease of use, that makes them accessible even to non-specialists. Combined with the ensemble methods, they furthermore often provide state-of-the-art results in terms of predictive accuracy. Common algorithms include the Classification And Regression Tree - CART (Breiman, 1984), the C4.5 (Quinlan, 1986) and the Random Forests (Breiman, 2001) methods. Importantly in the context of high-throughput data sets, tree-based methods are also highly scalable from a computational point of view. In its simplest form, a decision tree combines several binary tests in a tree structure. As an illustration, Figure 15 presents a bi-dimensional classification problem where goal of learning is to find a function which discriminates at best between red and green points.

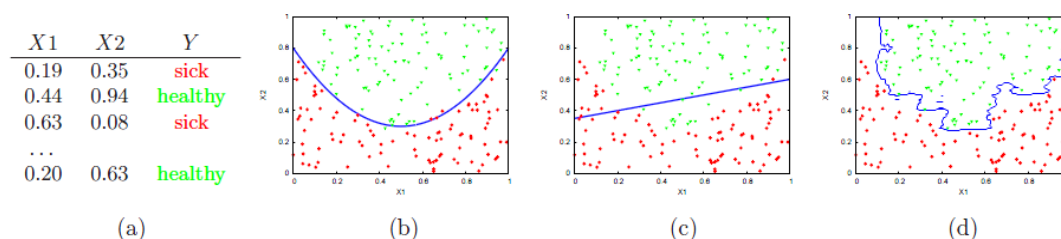



Figure 15: An illustrative two-dimensional supervised learning problem: From left to right: (a) tabular learning sample, (b) scatter-plot of the learning sample together with the optimal classification boundary for this problem, (c) the classification boundary for a too simple model and (d) of a too complex one

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	26/61

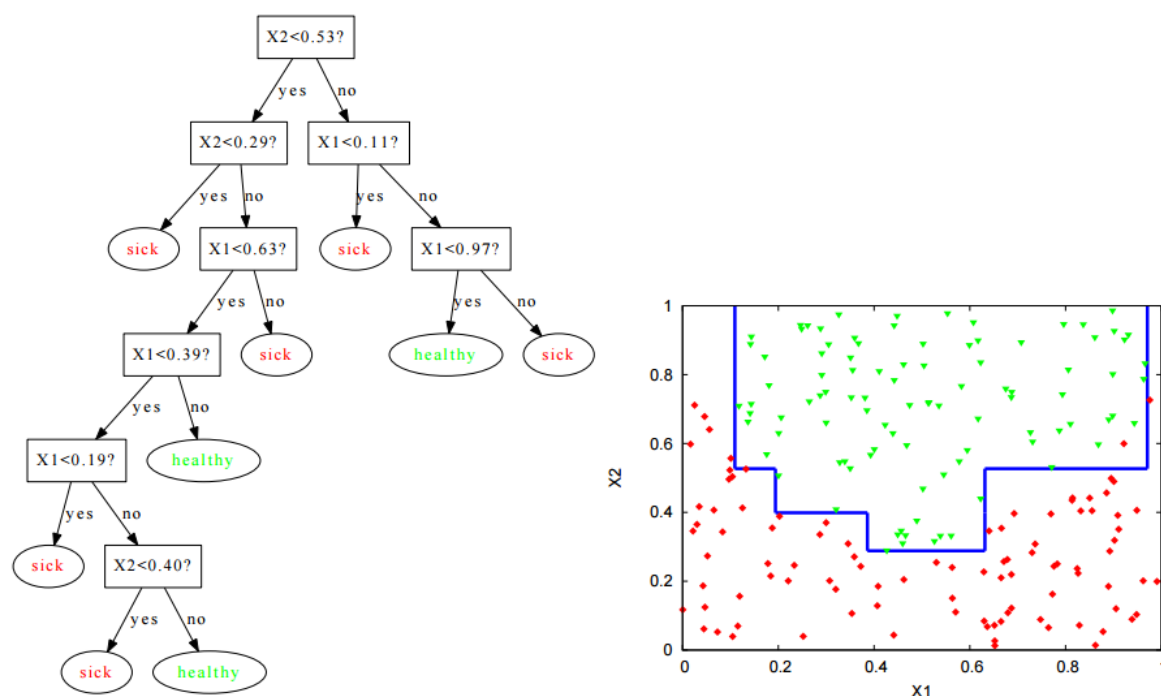



Figure 16: A decision tree and the corresponding decision boundary

For example, **Errore. L'origine riferimento non è stata trovata.** plots a decision tree and the resulting decision boundary. Each interior node of this tree is labeled with a test, which compares the value of an input attribute to a threshold, and each terminal node is labeled with a class. To produce a classification for a new object whose attribute values are known, we simply propagate it into the tree from the top node according to the test answers. When a terminal node is reached, its corresponding class label is attributed to the object. By using such tests, the tree progressively partitions the input space into hyper-rectangular regions where the output is constant. The general idea behind tree induction algorithms is to find a simple tree that has good predictive performance on the learning sample. Since the enumeration of all possible trees is essentially intractable, most tree induction algorithms are based on heuristics. The most common heuristic is a greedy top-down recursive partitioning approach. This algorithm starts with a single node tree corresponding to the complete learning sample and finds a way to split this node by selecting a test among a set of candidate tests. The algorithm then precedes recursively to split the successors of this node. The whole process results in a partition of the learning sample into smaller and smaller subsets. The development of a branch is stopped when some stop-splitting criterion applies. Eventually, each terminal node of the tree is labeled with a prediction (class name or vector of class probabilities) which is computed on the basis of the subset of objects which reach this node. This tree growing step is then usually followed by a pruning stage which aims at removing unessential parts of the tree, so as to avoid over-fitting. Possible application in HEALS include, the prediction of interactions between different types of bio-molecules, such as Protein-Protein Interactions (PPIs) and DNA protein interactions. It is noted that the prediction of the PPIs requires adjustment of the tree learning algorithms, because each attribute is measured twice, once for each protein of a pair.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	27/61

4.4.6 Artificial Neural Networks

Artificial neural network (ANN) are a family of statistical learning algorithms implemented in a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output. ANNs are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found. ANNs have three layers that are interconnected (Figure 17).

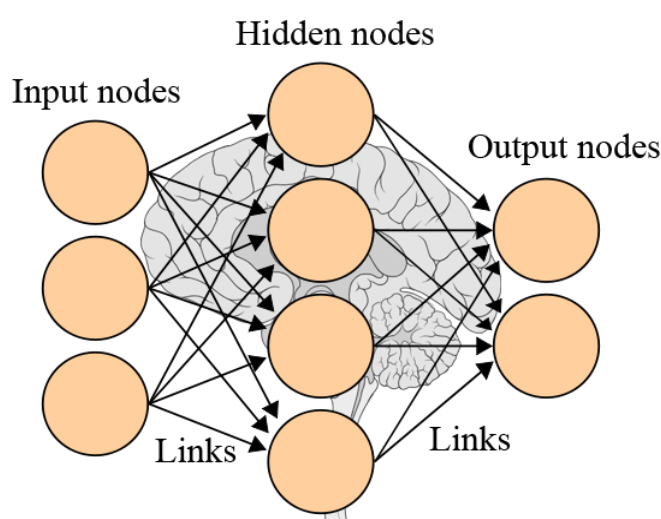



Figure 17: Typical ANN architecture

The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer. As ANNs are loosely based on the way a biological neuron is believed to organize and process information, they have many advantages in their ability to derive meaning from large complex datasets. Firstly, they do not rely on data to be normally distributed, an assumption of classical parametric analysis methods. They are able to process data containing complex (non-linear) relationships and interactions that are often too difficult or complex to interpret by conventional linear methods. Secondly, they are fault tolerant, i.e. they have the ability of handling noisy or fuzzy information, whilst also being able to endure data which is incomplete or contains missing values. Thirdly, they are capable of generalization (like other machine learning methods), so they can interpret information which is different to that of the training data, thus representing a 'real-world' solution to a given problem by their ability to predict future cases or trends based on what they have previously seen. Thus, trained ANNs can be used as standalone executable systems in order to predict the class of an unknown case of interest and therefore have the potential application in diagnosis. Finally, there are several techniques that can be used to extract knowledge from trained ANNs, and the importance of individual variables can be easily recovered using various methods such as the analysis of interconnecting network weights (Olden et al., 2004), sensitivity analysis and rule extraction (Silva et al., 2008). ANNs also have their limitations: training of ANNs can potentially be time consuming depending on the complexity of the data being modelled, and as the number of hidden layers required to capture the features of the data increases,

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	28/61

so does the time taken for training to complete. As such, only one or two hidden layers are commonly used. Over-fitting may be a problem in ANNs, which is a memorization of the training cases causing the network to perform poorly on future cases. The one major barrier which researchers usually associate with ANNs is that it is not always apparent how they reach a solution, and because of this they have been referred to as ‘black boxes’ (Duh et al., 1998; Smith et al., 2003; Tung et al., 2004; Wall et al., 2003). In addition, the quality of the model output is highly dependent upon the quality of the input data. If the input data is not representative of the ‘real world’ scenario, the model is compromised. To overcome these issues, several techniques for pre-processing the data have been proposed, and the reader is referred to (Barla et al., 2008; Phan et al., 2006; Wang, 2008; Wong et al., 2005a; Wong et al., 2005b) for more examples.

4.4.7 Support Vector Machines

Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Neural networks (NN) and Radial Basis Functions (RBFs), both popular data mining techniques, can be viewed as a special case of SVMs. SVMs perform well on the bioinformatics and bio-sequence analysis. Their introduction in the early 1990s led to an explosion of applications and deepening theoretical analysis that established SVM along with neural networks as one of the standard tools for machine learning and data mining. Figure 18 depicts the optimal hyperplane which has the largest distance from the nearest training data of any class (functional margin). It is noted that the larger the margin the lower the SVM generalization error.

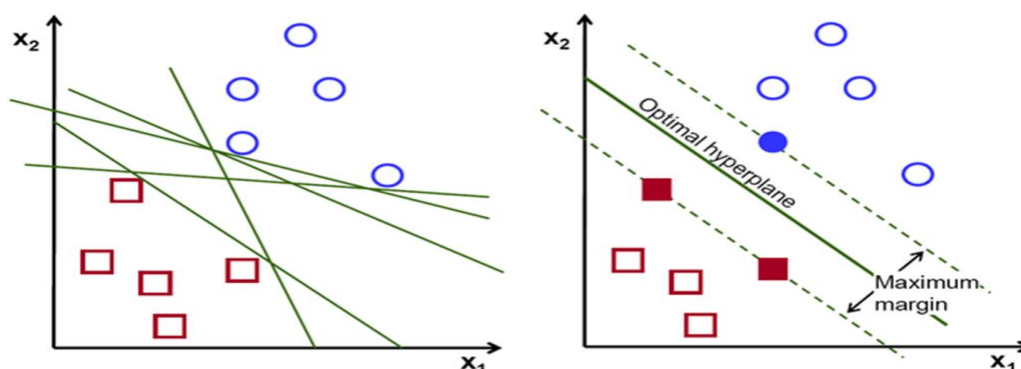



Figure 18: Detection of the optimal hyperplane and the maximum margin in clustered data

4.4.8 Bayesian Networks

Bayesian networks (Pearl 1988) represent the dependence structure between multiple interacting quantities (e.g., expression levels of different genes). Bayesian networks are a promising tool for analyzing for example the gene expression patterns. First, they are particularly useful for describing processes composed of locally interacting components; that is, the value of each component directly depends on the values of a relatively small number of components. Second, statistical foundations for learning Bayesian networks from observations, and computational algorithms to do so are well understood and have been used successfully in many applications. Finally, Bayesian Networks provide models of causal influence: Although Bayesian networks are mathematically defined strictly

 FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	29/61

in terms of probabilities and conditional independence statements, a connection can be made between this characterization and the notion of direct causal influence (Pearl, 1991; Spirtes, 1993). Probabilistic graphical models are graphs in which nodes represent random variables, and the (lack of) arcs represent conditional independence assumptions. Hence they provide a compact representation of joint probability distributions. Undirected graphical models, also called Markov Random Fields (MRFs) or Markov Networks, have a simple definition of independence: two (sets of) nodes A and B are conditionally independent given a third set, C, if all paths between the nodes in A and B are separated by a node in C. By contrast, directed graphical models also called Bayesian Networks or Belief Networks (BNs), have a more complicated notion of independence, which takes into account the directionality of the arcs, as we explain below. Undirected graphical models are more popular with the physics and vision communities, and directed models are more popular with the AI and statistics communities. (It is possible to have a model with both directed and undirected arcs, which is called a chain graph.) Although directed models have a more complicated notion of independence than undirected models, they do have several advantages. The most important is that one can regard an arc from A to B as indicating that A "causes" B. This can be used as a guide to construct the graph structure. In addition, directed models can encode deterministic relationships, and are easier to learn (fit to data). For a directed model, we must also specify the Conditional Probability Distribution (CPD) at each node i.e. $\Pr(V_i|\pi(V_i))$. If the variables are discrete, this can be represented as a table (CPT), which lists the probability that the child node takes on each of its different values for each combination of values of its parents $\pi(V_i)$, as seen in equation 2.

$$\Pr(V_1, \dots, V_n) = \prod_{i=1}^n \Pr(V_i|\pi(V_i)) \quad (2)$$

The advantages of the DAG models include that they handle efficiently incomplete datasets using novel sampling techniques (e.g. Gibbs sampling) and they can learn casual relationships for specific problem domains. It should also be noted that over fitting can be avoided by embedding the prior knowledge to the model structure. Figure 19 depicts the probabilistic inference, whereby it is possible to determine any probability of interest from an estimated GrAphical model (DAG) using a prior knowledge embedded to the model in order to avoid over-fitting.

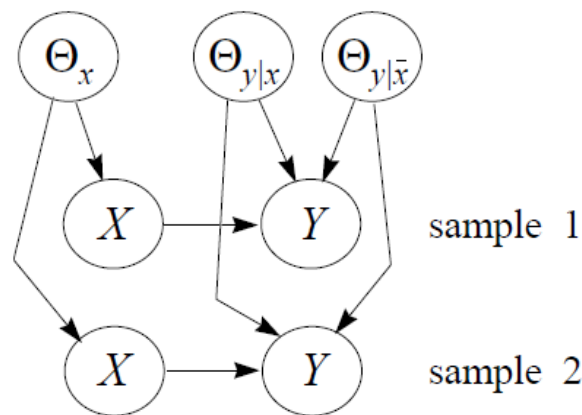



Figure 19: Implementation of the Bayesian networks

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	30/61

4.4.9 Fuzzy Logic


A fuzzy subset of X as a function of A i.e. $X \rightarrow [0, 1]$, is a characteristic function from X into the interval $[0, 1]$, where the $A(X)$ is called the membership function of the point x in the fuzzy set A , or the degree to which the point X belongs to the set A (Zadeh, 1965). The set of all fuzzy subsets (of X) is precisely the unit hypercube $I_n = [0, 1]^n$, as any fuzzy subset μ determines a point $P \in I_n$ given by $P = (\mu(x_1), \dots, \mu(x_n))$.

Reciprocally, any point $A = (a_1, \dots, a_n) \in I_n$, generates a fuzzy subset μ defined by $\mu(x_i) = a_i$, with $i = 1, \dots, n$. Non-fuzzy or crisp subsets of X are given by mappings $\mu : X \rightarrow [0, 1]$, and are located at the 2^n corners of the n -dimensional unit hypercube I_n .

Fuzzy logic has a number of successful application in area of bioinformatics, some listed here: to increase the flexibility of protein motifs (Chang, 2002), to study differences between polynucleotides (Torres, 2003), to analyze experimental expression data (Tomida, 2002) using fuzzy adaptive resonance theory, to align sequences based on a fuzzy recast of a dynamic programming algorithm (Schlosshauer, 2002), for DNA sequencing using genetic fuzzy systems (Cordón, 2004), to cluster genes from microarray data (Belacel, 2004), to predict proteins sub-cellular locations from their dipeptide composition (Huang, 2004) using fuzzy k-nearest neighbors algorithm, to simulate complex traits influenced by genes with fuzzy-valued effects in pedigreed populations (Carleos, 2003), to attribute cluster membership values to genes (Dembale, 2003) applying a fuzzy partitioning method, fuzzy C-means, to map specific sequence patterns to putative functional classes since evolutionary comparison leads to efficient functional characterization of hypothetical proteins (Heger, 2003), others used a fuzzy alignment model, to analyze gene expression data (Woolf 2000), to unravel functional and ancestral relationships between proteins via fuzzy alignment methods (Blankenbecler, 2003), or using a generalized radial basis function neural network architecture that generates fuzzy classification rules (Wang, 2003), to analyze the relationships between genes and decipher a genetic network (Agrawal, 1994), to process complementary deoxyribonucleic acid (cDNA) microarray images (Lukac. R., 2005) and to classify amino acid sequences into different super families (Bandyopadhyay, 2005).

4.5 Model Validation

Model validation techniques will be used to test the predicted model reliability using a number of well-known approaches including the Random Sub-sampling (RS), the K-Fold Cross Validation (KFCV) method, the Leave-One-Out Cross Validation method (LOOLV). Accordingly, using the Random Sub-sampling K data splits are evaluated, where each data split is randomly selected without further replacement. Alternatively, using the K-fold cross validation method involves partition to the dataset, where for each of the K experiments, $K-1$ fold are selected for training and the remaining fold for testing. The advantage of this method is that is that all examples in the dataset are eventually used for both training and testing. The Leave-One-Out Cross Validation method the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples. Here K experiments are performed and for each experiment $K-1$ examples are used for training and the remaining for testing. It is noted that for all the methods presented above the true error is estimated as the average error rate on the test examples. In addition other standard performance indices will be evaluated including the sensitivity, specificity and accuracy. Lastly the model performance can also be evaluated graphically, using the Receiver Operating Characteristic (ROC) metrics, depicting the steepness of the curve and the area under the curve.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	31/61

4.6 Model Analysis


The key aspects in any model analysis include: to visualize the model output, to map the underlying decision hyper-plane of the models, to depict the differences of the individual performance indices, to interpret the deduction mechanism. Certainly model interpretation is more complex for the case of the ANN and SVM models since it requires a better tuning of the proposed models architecture in order to optimize their predictions.


4.7 Meta-Modeling

The scope of the meta-learning techniques is to help select, combine or create optimal predictive models. Since there is no single method to outperform others, in accordance to the no-free lunch theorem, many different approaches should be selected since there is an optimal learning algorithm for each task. The choice of a meta-learning algorithm naturally depends on the problem and the task to be solved. Generally, traditional classification algorithms are very successful in meta-learning algorithm selection and can include meta-decision trees (Todorovski, 2003), neural networks, Support Vector Machines or any other classification algorithms, with the k-Nearest Neighbors being another popular choice (Brazdil, 2009; Brazdil, 2003). Applying regression algorithms is less popular, even smaller is the number of available algorithms to learn rankings. One of the simplest ranking method involves dividing the problem space using clustering of available datasets according to a distance measure (usually k-Nearest Neighbor) of the meta-features and using average performance ranks of the cluster into which a new problem falls. Other authors (Brazdil, 2003) also look at the magnitude and significance of the differences in performance. The NOEMON approach (Kalousis, 1999) builds classifiers for each pair of base forecasting methods with a ranking being generated using the classifiers' outputs. Build decision trees (Todorovski, 2002) use the positions in a ranking as target values but due to the complexity of the underlying HEALS datasets, we propose the combination of the different algorithms in order to reduce the probability of misclassification. In this regard we propose the use a meta-learning system (e.g. METAL project) for the automatic selection of the learning algorithms.

4.8 Biomarkers Fusion

The biomarker fusion combines data from different sources together, where the main objective of employing fusion is to produce a fused result that provides the most detailed and reliable information possible. Hence fuse multiple information sources together and produce a more efficient representation of the data. In accordance to *"task 7.3 of the model integration – biomarkers identification and prediction validation"* the prediction accuracy of these biomarkers will be tested against an independent test set, other than that ones used for training based on the principles of multivariate analysis. Hence, the biomarkers fusion will be realized efficiently through an inference system that is based on fuzzy logic. In addition since no prior knowledge exists in relation to the normal and pathological levels of the derived biomarkers, fuzzy logic rule sets will be considered to aid the design of robust decision support systems. Hence critical aspects about the biomarkers interoperability will be revealed which will lead to even better diagnostic procedure.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	32/61

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	33/61

5 Bioinformatics data infrastructure - data management and Integration


5.1 Introduction

The DIAMONDS (Data Infrastructure for Applying Models ON Design and Safety), available <https://diamonds.tno.nl/>, is a data infrastructure with statistical and computational tools aimed at predicting complex toxicological endpoints through integrated analysis. It includes *in vitro* effect-specific screening models for 'biological verification' for verification of *in silico*-based toxicity predictions, thus reducing the uncertainty often associated with *in silico* models for complex endpoints. The goal is to access the toxicological profile of a chemical molecule at an earlier stage and in a cost and a time efficient way. Within the HEALS context, DIAMONDS aims at providing the data infrastructure to support storage of data, metadata and analysis pipelines for omics data emerging from cohorts (DIAMONDS and GSCF/dbNP) and integrate this into the HEALS database platform developed in WP12. Integrated into DIAMONDS are tools available for quick omics-based pathway enrichment analysis (e.g. ToxProfiler) across multiple datasets from different studies. The DIAMONDS infrastructure will be used to support descriptive and predictive data mining outlined in the above tasks, on data streams from cohort studies captured in dbNP, to provide confirmatory support from public *in vitro* and *in vivo* animal genomics data related to chemical stressors.

5.2 Concept

5.2.1 The Phenotype Database (dbNP)

Within HEALS, data infrastructure support is needed to store and analyse omics data obtained from human cohorts. The Phenotype Database (dbNP) (<http://phenotypefoundation.org/>) is a bioinformatics application that can store any biological study. It contains templates which makes it possible to customize. The main module of dbNP is the Generic Study Capture Framework (GSCF). In order to allow flexibility to capture all information required within a study, and to make it possible to compare studies or study data, the system uses customizable study design templates and ontologies (Figure 20). It is especially designed to store complex human study designs including cross-over designs and challenges. In addition, it contains a Transcriptomics module, a Metagenomics module, a Metabolomics module and Simple assay module, which allows for the analysis and data integration with cohort specific metadata. Phenotype Database facilitates sharing of data within a research group or consortium, as the study owner can decide who can view or access the data. New studies can be based on study data within the database, as standardized storage is stimulated by the system. As such, it represents an important data infrastructure component for omics studies applied in human cohort settings as foreseen in HEALS. The dbNP GSCF will be customized to accept HEALS cohort data and study designs, together with associated omics and assay data. In addition, connections concerning data export will be established with the larger HEALS infrastructure Database platform developed in WP12.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	34/61

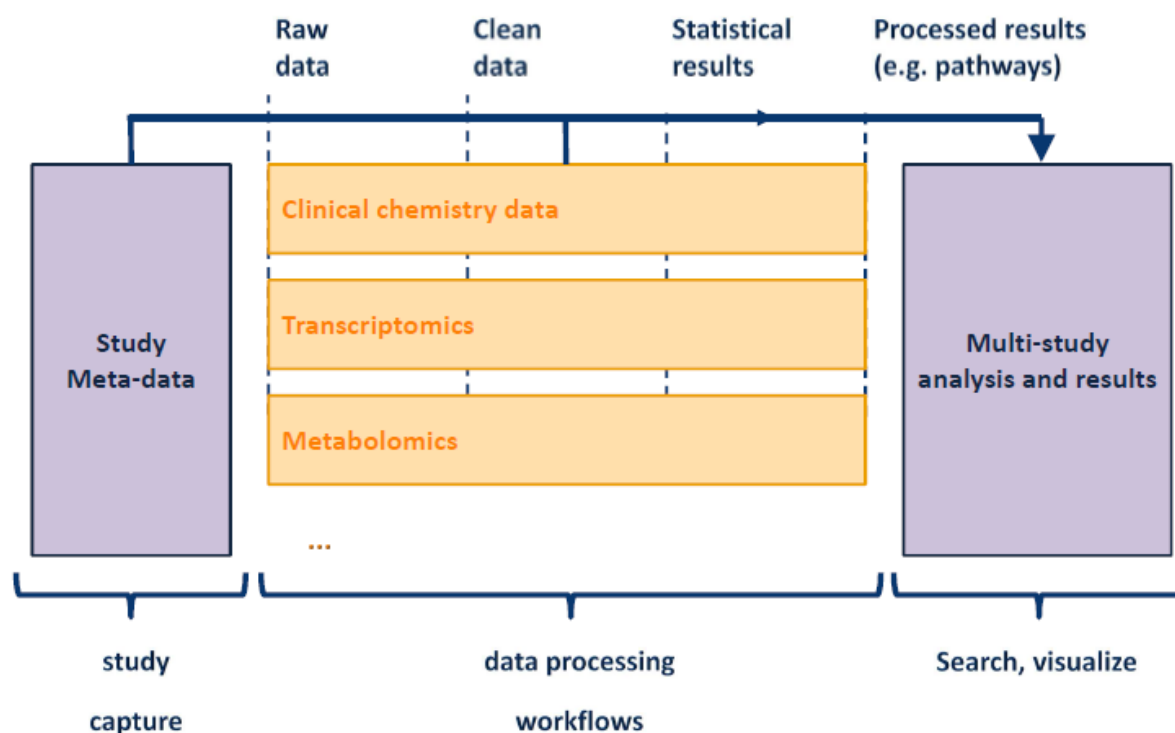



Figure 20: Collect, harmonize and distribute data

5.2.2 Data Infrastructure for Applying Models ON Design and Safety (Diamonds)

DIAMONDS captures data from diverse domains, i.e. chemical structure information, *in vitro* (HTS) assay data, *in vivo* and *in vitro* omics data and *in vivo* clinical chemistry and pathology. Furthermore, it allows the user to import public data besides private or consortium-derived data and handles access on group level. DIAMONDS also connects using application programming interfaces (API's) to diverse knowledge databases and bring together compound structure information, pathway/geneset information, study metadata information and toxicological concepts. In addition, knowledge is matched to existing ontologies and users can easily link through to external databases. The DIAMONDS infrastructure is built upon a combination of a Unix file server, backup server, calculation server, MySQL database and a website based on a combination of HTML, PHP and javascript. From the website, the data are brought together in diverse views with main modules centered around structures, compounds, studies, meta-analysis, classification, pipelines and users. Data can be examined using diverse visualization and analysis tools, such as QC reports, basic multivariate data analysis, principal component analysis, similarities, classification and pathway scoring. In addition, DIAMONDS is designed to allow pathway based meta-analysis by integrating data together with pathway scores generated by the integrated ToxProfiler module. Figure 21 illustrates the main aspects of the DIAMONDS infrastructure, including the knowledge base tool, the external source information and the analysis tools.

 FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	35/61

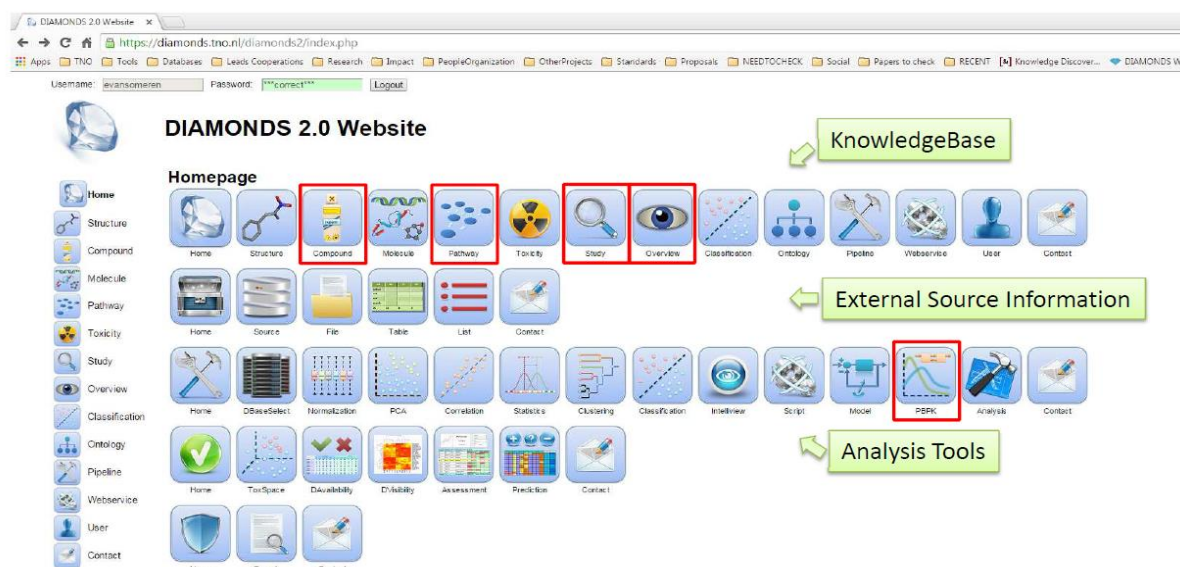


Figure 21: The DIAMONDS 2.0 website and its main components

In addition, Figure 22 illustrates how researchers can visually inspect structural features of compounds and see which compounds are involved in which pathways and used in which studies. In this way an NTC researcher knows directly which studies might be of relevance to perform a meta-analysis on. In addition, structural similarity is based on 2D or 3D fingerprints or more structurally enriched fingerprints, commonly used in pharma, such as ECFP4 and FCFP4. Furthermore, when viewing a compound, DIAMONDS provides information on whether this compound is used in one of the available AOPs such that its known mechanism can be taken into account when interpreting data.

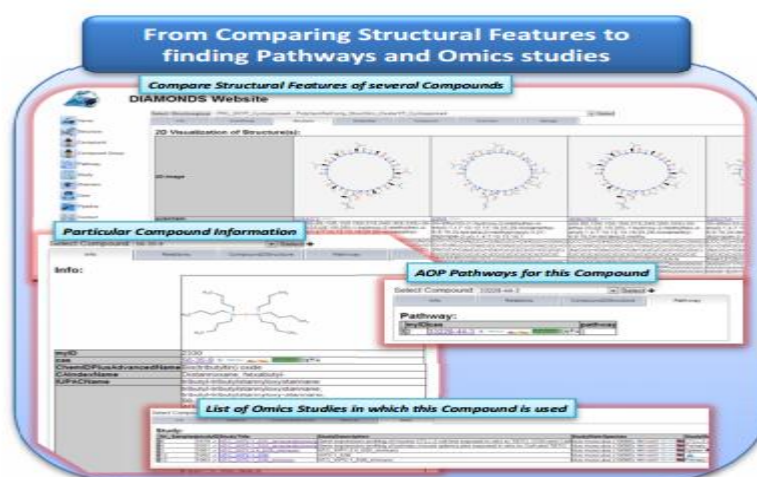



Figure 22: DIAMONDS links structural features of compounds to studies and pathways

Figure 23 illustrates with screenshots how metadata is captured on study-level and sample-level. If possible, controlled vocabularies and/or ontologies are being employed and link-outs to the external references, such as BioPortal, allow users to understand the terms used. However, as ontologies are

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	36/61

never complete, users are allowed to add new terms in case no fitting term can be found (e.g. some cell-lines are difficult to find in ontologies). For sustainability, DIAMONDS allows metadata to be saved as ISATAB, to be used in the ISATAB environment. Furthermore, an API between DIAMONDS and the Nutritional Phenotype database (dbXP) allows studies stored in dbXP to be shown in DIAMONDS as well. Normalization and Quality Control is of utmost importance when handling Omics data. Arrayanalysis.org is a free online tool that handles normalization of transcriptomics data. DIAMONDS integrates with the output of Arrayanalysis.org and allows users to visually inspect the QC-results (MAplots, heatmaps, etc). As metadata is stored, it is easy to generate a PCA to inspect the structure inside the data and change labeling and coloring according to available sample annotations.

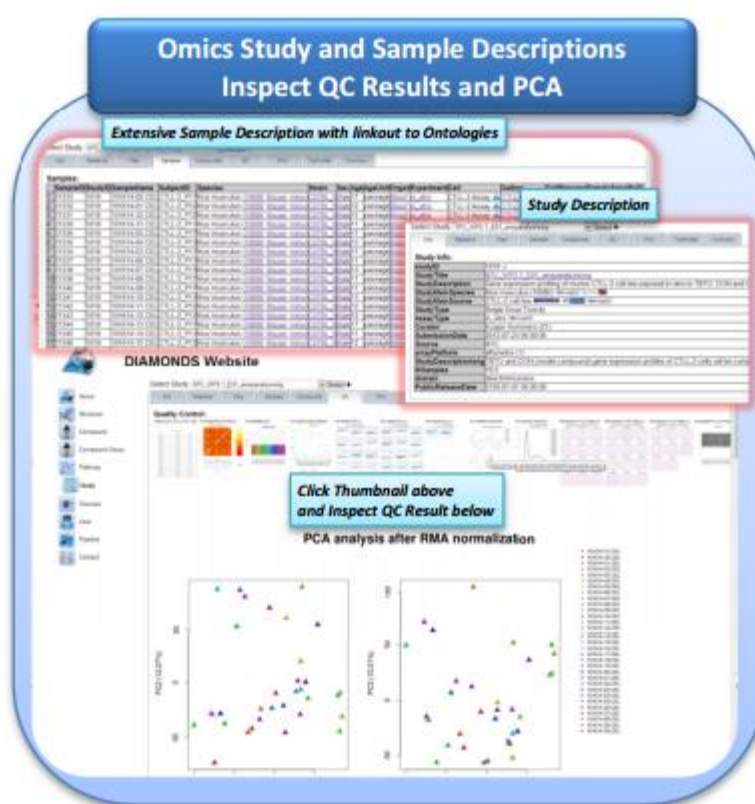



Figure 23: Screenshots of DIAMONDS illustrating capture of metadata and QC results

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	37/61

6 Implementation of big data analytics – GeneSpring

6.1 General characteristics

An efficient way to implement the big data analytics required in the bioinformatics analysis for HEALS is the adaptation and use of dedicated software such as GeneSpring by Agilent Technologies. The latter provides the necessary statistical tools for intuitive data analysis and visualization. Its interactive environment allows the functional integration of Transcriptomics, Metabolomics, Proteomics and NGS data within a biological context. Heterogeneous data that measure various biological entities and events such as mRNA and microRNA expression, exon splicing, DNA structural variation, proteins and metabolites will be analyzed. In addition, the metadata analysis and available visualization tools will allow us to analyze the phenotypic parameters such as the clinical or physiological attributes of the subjects alongside their gene or metabolite expression profiles. Complementing more traditional bioinformatics techniques, correlation tools and visualizations we will identify the co-regulated genes, the metabolites, and the proteins in an intuitive and easy-to-use manner. The prior information will be utilized in designing follow-up experiments, using next-phase experiments from pathway information, enable us with hypothesis-driven experimental design via the incorporation of prior biological knowledge from multiple measurement technologies according to the HEALS high dimensional biology paradigm.

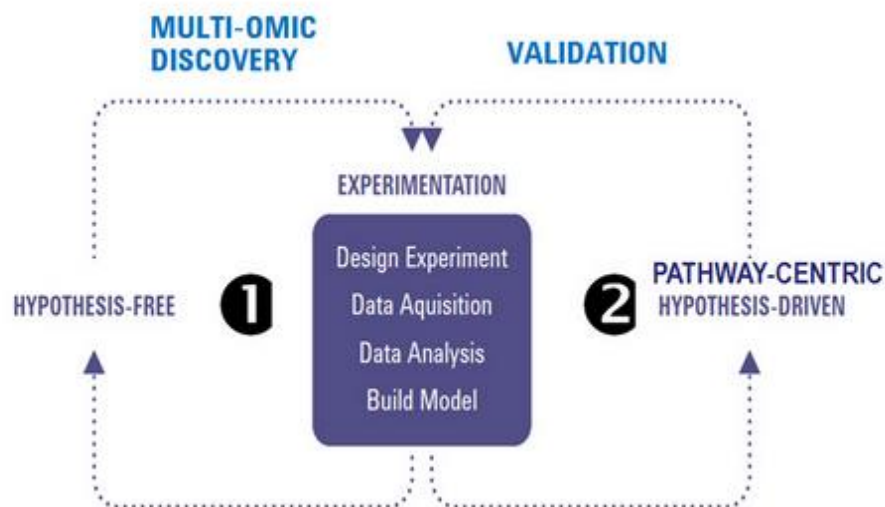



Figure 24: Functional flowchart of Genespring

Figure 24 presents a GeneSpring flowchart of both the multi-omic discovery and the validation section. It is noted that heterogeneous data such as gene expression, miRNA, exon splicing, genomic copy number, genotyping, proteins, and metabolites, will be combined into one bioinformatics project, allowing us to analyze, compare, and view results as seen below.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	38/61

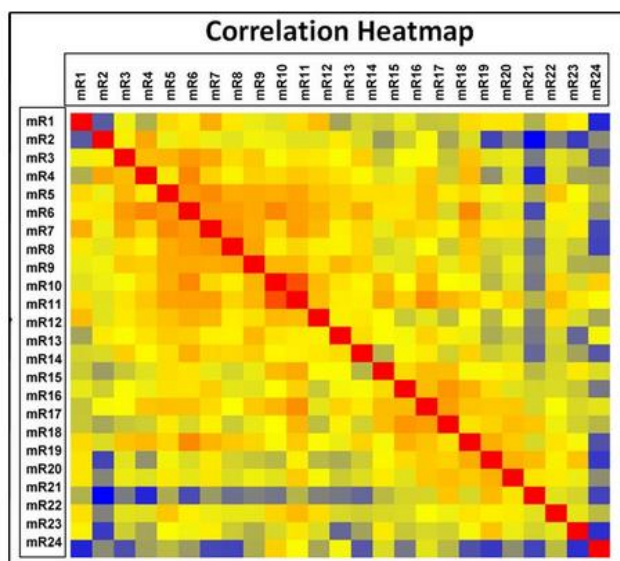



Figure 25: Heatmap with correlation coefficients between entities of a single experiment

As shown in Figure 25, correlation analysis will allow for the identification of the co-regulated molecules such as gene and metabolite and to identify relationships between the samples. The correlation framework is supported on most of the datasets generated using high throughput omic platforms such as Microarray, Mass Spectrometry and Next Generation Sequencing. Furthermore, the framework supports pair-wise correlations, using a single technology platform and cross-technology measurements between two different platforms. Hence correlation analysis will be performed in a pair-wise manner in order to identify and visualize dependency between abundance levels of any pair of biological entities. It is noted that an “entity” in GeneSpring is defined as a gene, metabolite, protein or a probe in an expression array. It is this semantic flexibility that led us to use GeneSpring as the data integration platform of choice in HEALS. Correlation analysis can be performed on entities within a single experiment or across two different experiments. For cross-experiment correlation analysis a Multi-Omics Analysis (MOA) experiment is created in GeneSpring using the two experiments whose entities would be selected for correlation. Table 1 summarizes the types of experiments, which support correlation analysis.

Table 1: Experiment Types in GeneSpring Available for Correlation Analysis

Experiment Type	Within Single Experiment	Across Two Experiment
mRNA Expression	YES	YES
Exon Expression	YES	YES
miRNA	YES	YES
RTPCR	YES	YES
DNA-Seq	NO	NO
RNA-Seq	NO	YES*
smallRNA-Seq	NO	YES*
Metabolomics	YES	YES
Proteomics	YES	YES

In addition to entity-entity correlation, pair-wise correlation analysis between biological samples within a given experiment is also supported. Sample correlation will allow us to identify the condition-wise relationships which may exist between the samples in a study, as seen in Figure 26.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	39/61

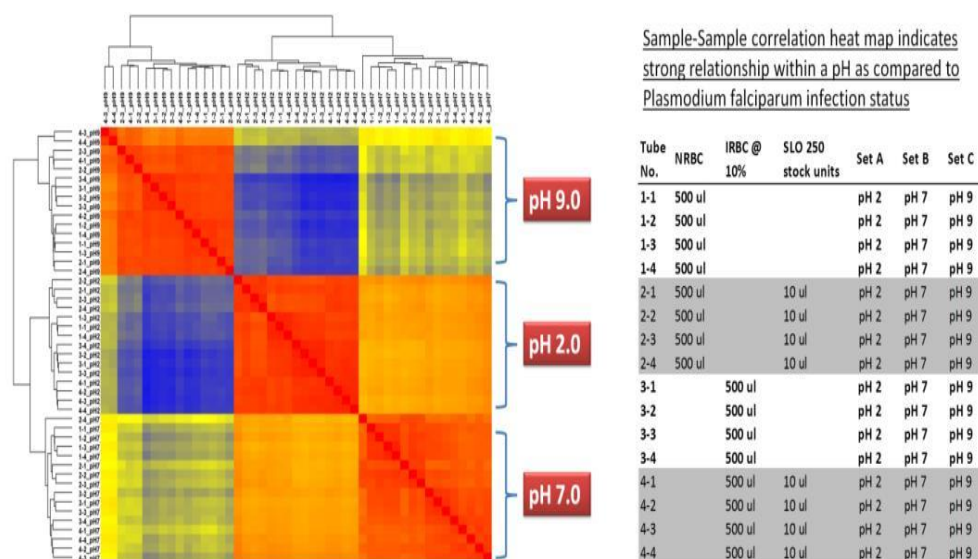


Figure 26: Sample-sample correlation heatmap for a metabolomics study. Clustering on correlation coefficients clearly demonstrates that samples group based on their pH values rather than the infection status (NRBC- Non-infected RBCs; IRBC-Infected RBCs).

6.2 Multi-omic Pathway Analysis

To further elucidate the context of pathways and interaction networks, multi-omic pathway analysis is essential. This would facilitate insights into the underlying biology. Pathways are imported and viewed in the KEGG, Wikipathways (GPML format), BioCyc and BioPAX exchange format. These pathways will be used in single experiment analysis and the Multi-omic analysis pathway tool, which could aid us to determine if there is an enrichment of the entities of interest in any pathways, seen in Figure 27.

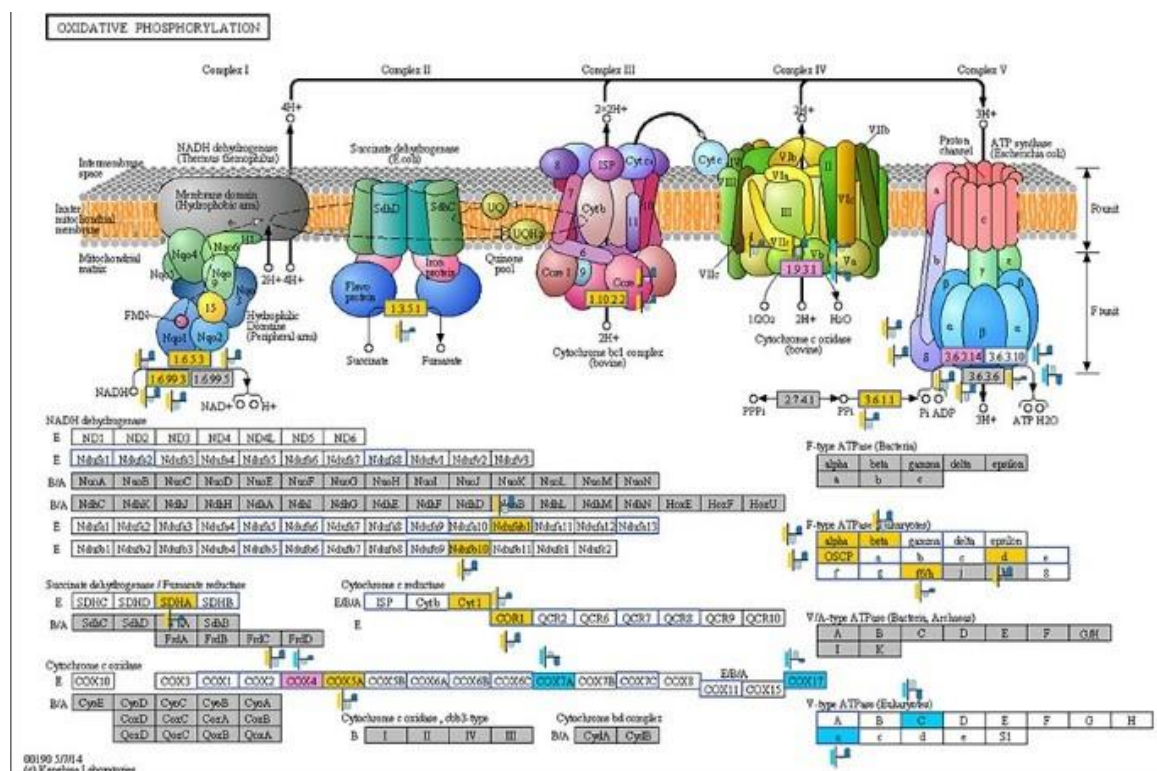



Figure 27: Pathway and network diagrams help place statistical results in a biological context. Direct navigation between biological pathways and their associated genes provides systems-level insight.

Pathway Analysis and Visualization options allow us to do the following:

- Identify of significant pathways from multi-omic data
- Canonical biochemical, metabolomic, & signaling pathways from KEGG, Wikipathways and BioCyc
- Curated pathway rendering
- Intuitive data overlay
- Support for GPML / OWL pathway import
- Custom pathway creation
- Pathway browsing, searching and navigation
- Automatic translation between annotation types, pathways, and organisms

6.3 NLP Network Analysis

Since genes and proteins interact in a biochemical network to orchestrate the biological processes involved in a disease, it is useful to be able to generate and dynamically explore such networks. For example, using a set of algorithms and provided organism-specific interaction databases, this could help us build a range of network types, including targets and regulators, transcription regulators,

 FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	41/61

biological processes, and shortest connect. Natural language processing-based (NLP) algorithms can be applied to a body of text, HTML, a PDF, or Medline XML to extract and add interactions to an existing interaction database (Figure 28).

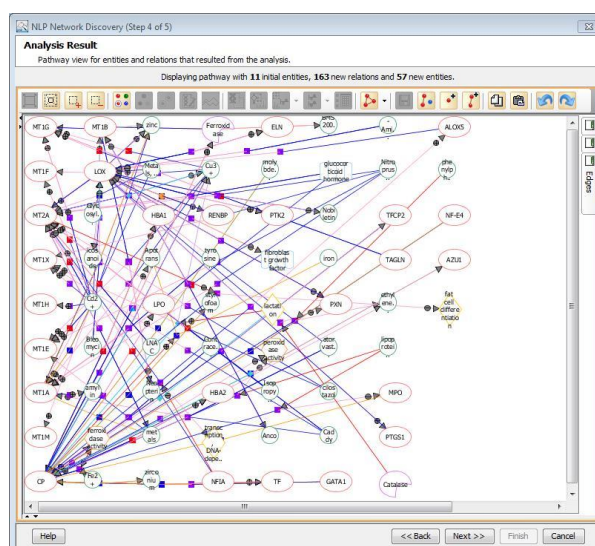



Figure 28: NLP (Natural Language Processing) Discovery Network

In addition we could extend the biological contextualization capabilities through integration with Ingenuity Pathway Analysis (IPA), Metacore and CytoScape, where lists of genes and experimental data could be seamlessly transferred between the two applications for iterative analysis.

6.4 Meta-data framework

Clustering analysis is an efficient way to group the samples and conditions in a dataset into subsets based on the similarity of their abundance profiles (Figure 29 and Figure 30). Sample clustering has been broadly used for inferring disease subtypes and for patient stratification. Used in this context, the hierarchical clustering is a very important analysis tool for revealing the molecular mechanism underlying the biological function. The software will allow us to perform a metadata analysis and allows to visualize the abundance profiles of samples alongside metadata such as clinical, physiological, or technology related information. The metadata visualization framework could help us reveal tacit dependencies between characteristics of the subjects or samples and their gene, metabolite, or protein expression profiles.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	42/61

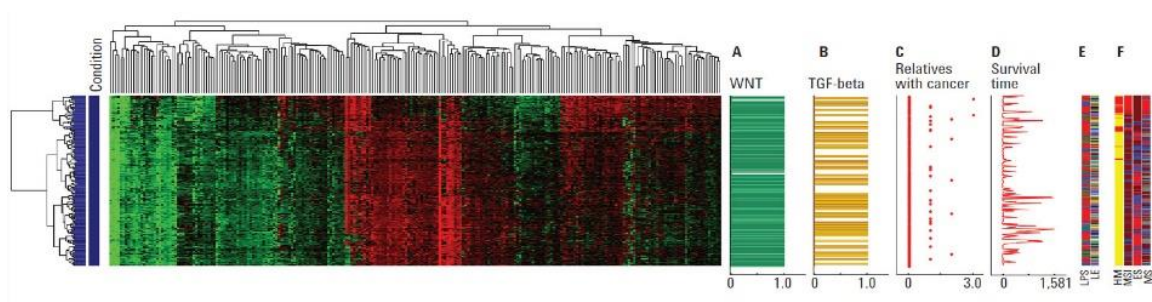


Figure 29: Hierarchical clustering of 220 colon cancer samples from TCGA with the visual alignment of metadata. (Labels in panels E and F: LPS – Lymphnode pathologic spread, LE – Lymphnode examined, HM-Hypermuted, MSI – Microsatellite Instability Status, ES – Expression subtype, MS – Methylation subtype).

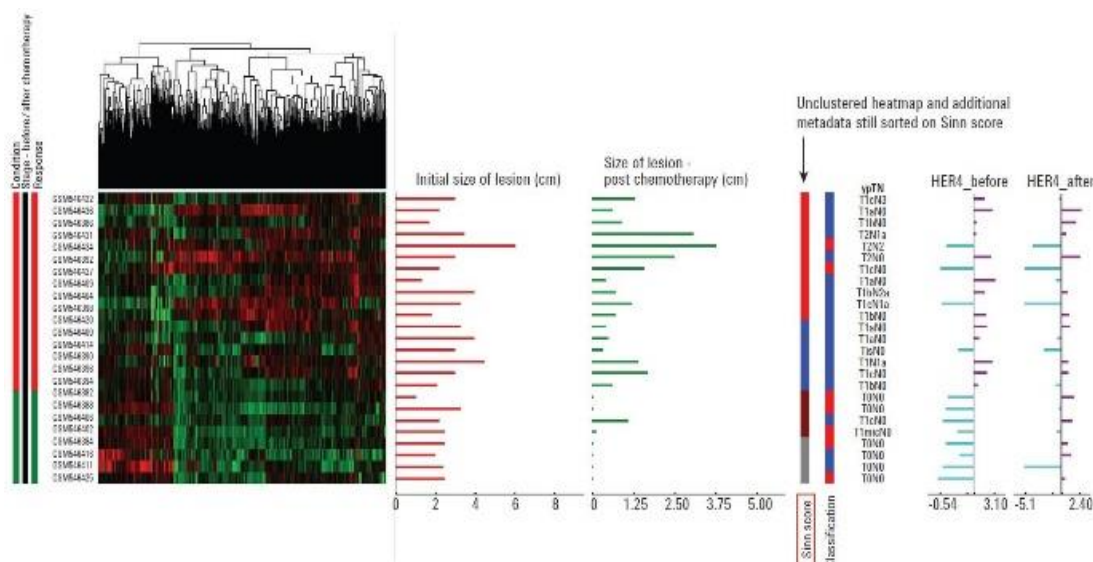



Figure 30: HER4 shows up-regulation in patients with tumor regression (Sinn score 3 and 4). Normalized expression values for HER4 (probe A_32_P183765) are shown before and after chemotherapy.

6.5 Transcriptomic analysis

Flexible and comprehensive workflows for a variety of transcriptomic applications, including a broad spectrum of data pre-processing, linear and non-linear normalization methods for both one- and two-color gene expression data will be available (Figure 31). Hence data analysis can be performed using either the 'Guided Workflow' or 'Advanced Analysis mode'. Quality control can be performed using platform-specific metrics, so as to optimize pre-processing steps before statistical analysis.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	43/61

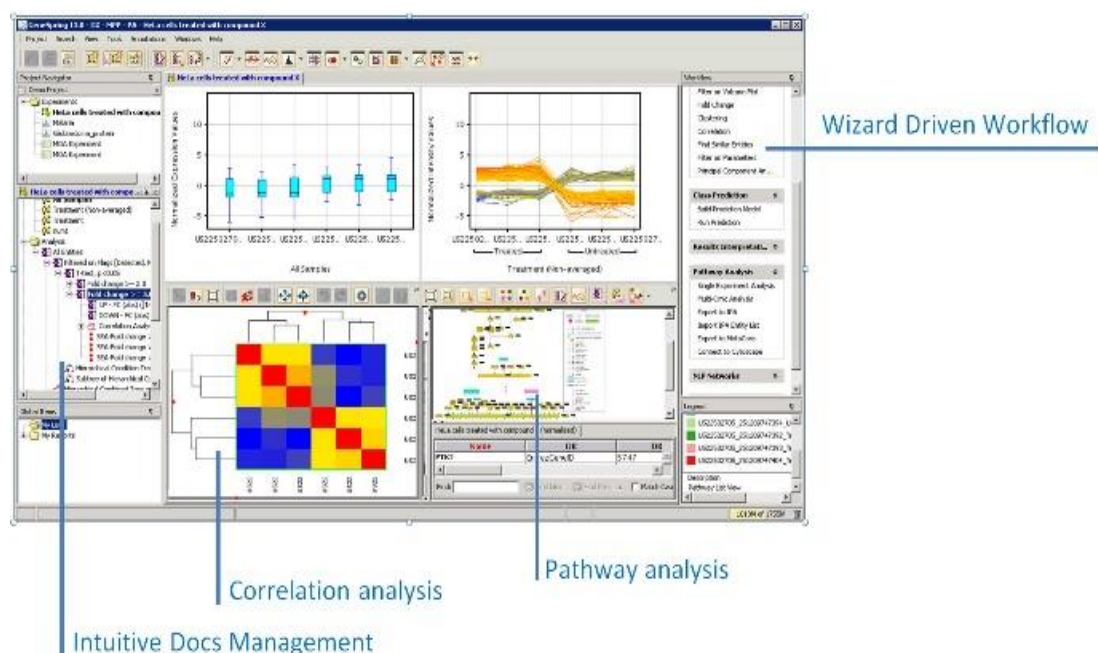



Figure 31: Intuitive GUI allows users to toggle between windows.

Other key features for Transcriptomics applications include:

- Probe-or gene-level expression analysis on all major microarray platforms, including Agilent, Affymetrix, and Illumina
- microRNA analysis and identification of gene targets using integrated TargetScan information
- The ability to do correlative analysis on mRNA expression and miRNA data, (or splicing, QPCR, copy number, etc.)
- Exon splicing analysis using t-tests or multivariate splicing ANOVA and filtering for transcripts on splicing index
- Visualization splicing analysis
- Real-time PCR QC and data analysis
- NCBI Gene Expression Omnibus Importer tool for expression datasets

6.6 Genomic copy number analysis

The interrogation of genomic structural variations and their implications in disease susceptibility and progression is supported as well. Providing workflows for paired and unpaired analysis of Affymetrix and Illumina genotyping array data, quick detection of regions of genomic copy number variation (CNV) and loss of heterozygosity (LOH) is supported. Once regions of interest are discovered, genes overlapping those regions can be identified and the biological impact of copy number variation can

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	44/61

be assessed in downstream GO Analysis or pathway analysis. Features of the genomic copy number workflow include:

- Ability to create and use a custom reference in addition to packaged HapMap reference
- Batch effect correction method
- Circular Binary Segmentation
- Filters to identify copy-neutral LOH events and regions of allelic imbalance
- Ability to identify common variations across a set of samples


6.7 Genome-wide association analysis

Genome-wide association studies (GWAS) utilize high-density genotyping microarrays to identify SNPs associated with qualitative or quantitative traits. The flexible workflow supports case-control experimental designs, offering a suite of statistical tests applied under various genetic models, multiple testing correction, and correction methods for population stratification. After identifying genes harboring SNPs or haplotype blocks associated with trait, GO Analysis and pathway analysis can be performed to determine what biological process and pathways may be involved in the disease under study. Other key features for the genetic association workflow include:

- EIGENSTRAT and Genomic Control population stratification correction
- Tag SNP identification
- Haplotype inference and Haplotype Trend Regression
- Pearson's Chi-Square, Fisher's Exact, Cochran Armitage, and Chi-Square correlation
- Logistic and linear regression
- LD plot

6.8 Statistical tools for testing differential expression

Clustering algorithms can be employed to group entities and/or samples based on the similarity of their expression profiles, revealing information regarding the biological function or the co-regulation of genes. Robust classification algorithms that use training datasets to find clinically predictive expression patterns. By offering multiple classifiers including Decision Tree, Support Vector Machine, Naive Bayesian, Neural Network, and Partial least squares discriminate, GeneSpring enables biomarker discovery for a variety of experimental designs.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	45/61

6.9 Extensible functionality with Jython and R

GeneSpring provides the ability to write, execute, and save custom scripts to combine operations with a more general Jython (Python with Java class import capabilities) programming framework. Hence we could develop our own data transformation operations, automatically pull up data views, and run external algorithms within the software. An embedded R scripting editor allows R scripts to be written and run from within the GeneSpring software. In addition any R function can be given access to GeneSpring data, with results being automatically incorporated back into the platform.

GeneSpring R integration:


- GeneSpring R package for:
- Reading and downloading GeneSpring datasets into R
- Creating new GeneSpring experiments from R
- Accessing Samples, Attributes, and Sample Attachments
- Accessing Experiment Grouping Information
- Accessing and Creating GeneSpring Entity Lists
- Searching on Projects, Experiments, Entity Lists, Technologies, and Samples

6.10 Report Generation Capability

Plot views and pathways including matched pathway lists can be exported as a Report. A report can be downloaded as a .pdf file or saved locally or on the cloud as a .gsreport file. Reports can be created to include:

- Legend
- Annotation columns to include experiment data
- Customized paper size
- Customized page orientation
- Customized page margins

Multiple reports can be merged to create a single report (Figure 32). For a report, the maximum number of pages to be displayed in the platform can be modified from Tools - Options. If a report contains more than the specified number of pages, it can be downloaded as a .pdf to the local system.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	46/61

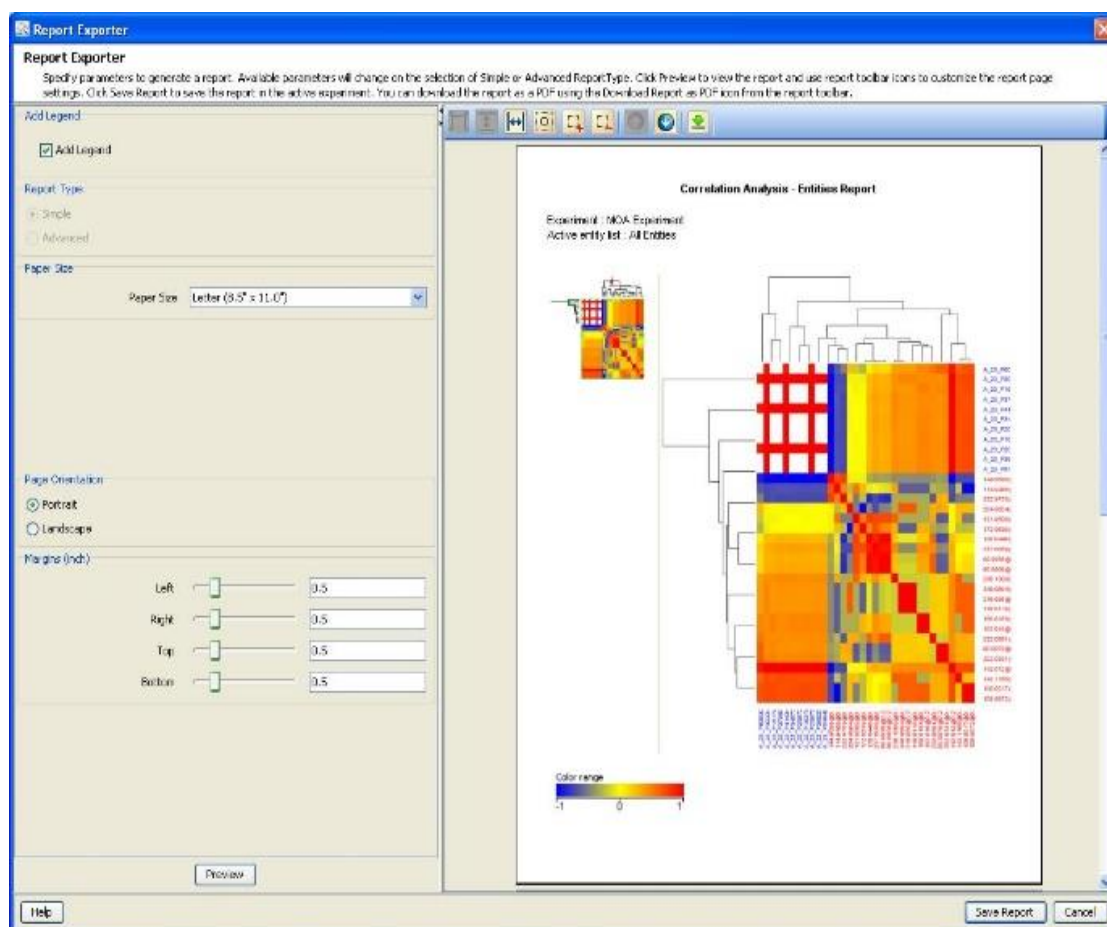



Figure 32: Report Generation Feature – it enables users to create pdf reports on the fly containing text and images.

6.11 Modules and file formats

To conclude, GeneSpring contains 4 modules, the GeneSpring GX, the GeneSpring MPP, the Pathway analysis and the Strand NGS workflow. The contents of each module are presented in Figure 33.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	47/61


GeneSpring GX	GeneSpring MPP	Pathway Analysis	^{new} GeneSpring – Strand NGS workflow
<ul style="list-style-type: none"> mRNA miRNA Alternative Splicing Copy Number GWAS qPCR Custom Arrays Correlation Analysis Metadata Framework for clustering view Report Generation 	<ul style="list-style-type: none"> Proteomics Metabolomics Support for GC/MS, LC/MS, CE/MS and ICP-MS Support for NMR data Generic Import Integrated ID Browser Correlation Analysis Metadata Framework for clustering view Report Generation 	<ul style="list-style-type: none"> Multi-omic Pathway Analysis Network Discovery Natural Language Processing MeSH Networks Support for KEGG pathways BioCyc, Wikipathways, BioPax Level III BridgeDb Mapping Support for External Pathway Tools 	<ul style="list-style-type: none"> Export of data objects from Strand NGS 2.1 and import in GeneSpring Data visualization Correlation analysis Pathway analysis

Figure 33: GeneSpring modules

In addition, the major file formats per GeneSpring module are tabulated in table 2, where the input file formats are presented. It is noted that the output file formats can be exported either in a pdf form (i.e. the pdf report) or via Jython programming in any common file format (i.e. txt or xls).

Table 2. GeneSpring module and corresponding file formats


Module	File format
Affymetrix Expression	.CEL / .CHP
Affymetrix Exon Expression	.CEL / .CHP
Affymetrix Splicing	.CEL / .CHP
Experiment Type	.CEL / .CHP
Affymetrix copy Number	.CEL / .CHP
Affymetrix Association analysis	.CEL / .CHP
Illumina Single Color	.txt
Illumina Copy Number	.txt
Illumina Association Analysis	.txt
Agilent Single Color	.txt
Agilent Two Color	.txt
Agilent miRNA Experiment Type	.txt
Pathway Experiment	.txt, .tsv, .csv and .xls
RT-PCR Experiment Type	- RQ 1.2, RQ 2.1, RQ 2.2 and RQ 2.3 formats of the ABI's 7900HT RT-PCR system
Generic Single Color	.txt, .tsv, .csv and .xls
Generic Two Color Experiment Type	.txt, .tsv, .csv and .xls

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	48/61

7 Application of big data analytics - Exposure to real life ambient air mixture


A very comprehensive example of applying omics analysis for understanding the environmental effects on public health is the emerging issue of exposure to multiple stressors. Toxicity of chemical mixtures is only partially addressed by the current state of environmental health sciences. The so-called 'cocktail' effect becomes all the more complex due to the large number of possible combinations of chemicals and other (physical or biological) stressors in the environment. This has hampered the development of rigorous methodologies for tackling the issue of environmental mixture safety. Currently, our scientific understanding and policy for environmental mixtures are based largely on extrapolating from, and combining, data in the observable range of single chemical toxicity to lower environmental concentrations and composition - i.e. using higher dose data to extrapolate and predict lower dose toxicity. More precise approaches to characterize toxicity of mixtures are needed. A major obstacle to the development of effective mixture risk assessment methodologies is the possibly infinite ways of combining chemicals into actual environmental mixtures (Mason et al., 2007). It should be noted, however, that although in theory the number of combinations of chemicals or stressors is infinite, the number of biological processes is finite. Therefore, in considering an integrated approach for risk assessment, it makes more sense to work on the finite biological processes that may be affected by human exposure to these mixtures rather than the infinite combinations of chemicals and stressors (Liao et al., 2002). Integrated health impact assessment of environmental stressor mixtures would need to follow a 'full chain' approach to take into account all relevant health stressors and their interaction. Application of the full chain approach entails considering all possible exposure pathways via the environment and lifestyle choices. It also encompasses considering the effects of co-exposure to relevant stressors and how risk modifying factors such as age, diet, gender, and time window of exposure affects the final physiological response. It is clear that successful application of this approach poses demanding data requirements both in terms of environmental monitoring and in terms of biological and clinical data interpretation. What is most important is the need for comprehensive data interpretation of the molecular, biochemical and physiological processes that couple exposure to health outcome.

This requires forging a new paradigm for interdisciplinary scientific work in the area of environment and health. We shall call this the connectivity paradigm for chemical risk assessment, denoting an approach that builds on the exploration of the interconnections between the co-existence of multiple stressors and the different scales of biological organization explicitly described by omics that together produce the final adverse health effect (Workman et al., 2006). Connectivity marks a clear departure from the conventional paradigm, which seeks to shed light on the identification of singular cause-effect relationships between stressors and health outcomes. It entails creating a new way of combining health-relevant information coming from different disciplines, including (but not limited to) environmental science, epidemiology, toxicology, physiology, molecular biology, biochemistry, mathematics and computer science (Kitano, 2002a; Kitano, 2002b). The integration of these different information classes into a unique framework to better inform and support public health impact assessment of chemical mixtures in the environment could serve as a good example for the integration of the omics in the risk assessment process and the evaluation of population health impact assessment.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	49/61

In the context of applying systems biology approaches to chemical mixtures risk assessment, we move forward towards the definition of a co-exposure biology using optimally physiology and systems biology based modeling and data mining and assimilation algorithms. The aim is to couple the systems biology approach to the gene-environment interactions with the corresponding “physiome” approach. The variable levels of biological organization involved in this holistic view of mixture toxicology and cumulative exposure and risk assessment suggest that different technologies need to be brought to bear in order to obtain a comprehensive view of how co-exposure to multiple chemicals affects the overall phenotypic response of individuals. Technological variability introduces the need for better data integration and assimilation and for the development of novel data analysis and hypothesis generation and testing procedures in order to best elucidate the biological mechanisms underlying mixture toxicity. On the basis of the data currently available, the best available techniques, and the need for data integration, a tiered approach has been developed and outlined herein to support the connectivity approach to the assessment of health risk associated to combined exposure to multiple chemical stressors (Sarigiannis et al., 2009). This exposure biology approach to mechanism-based risk assessment of environmental chemical mixtures can be tackled with an integrated, multi-layer computational methodology, ideally comprising the following steps:


- a) Characterization of exposure factors quantifying the parameters that affect human exposure to environmental chemicals, such as time-activity relationships, seasonal and climatic variation, and consumer choice. These exposure factors can be used to derive aggregate and cumulative exposure models, leading in probabilistic exposure assessments. Aggregation can be done across exposure pathways and routes and even across different exposure scenarios, if the relevant exposure metric or the imputable biological or physiological effect can be related to these scenarios. For instance, exposure to volatile organic compounds (VOCs) such as benzene or toluene and mixtures thereof may occur both from environmental media and in specific occupational settings. A cumulative exposure scenario for these substances would have to take stock of the actual variability of exposure across these different settings throughout typical days for the same period in an individual’s lifespan.
- b) Current toxicological state of the art combines estimations of biologically effective dose with early biological events to derive dose-effect models, which can be used in combination with the probabilistic exposure estimates to derive biomarkers of exposure and/or effect. Combined use of epidemiological, clinical and genetic analysis data may shed light on the effect of risk modifying factors such as lifestyle choices and DNA polymorphisms. Observation of real clinical data and/or results of biomonitoring, if coupled with the exposure/effect biomarker discovery systems, can produce biomarkers of individual susceptibility and thus allow estimations of individual response to toxic insults. Toxicogenomics, comprising transcriptomics, proteomics and metabolomics, and adductomics (considering adducts of xenobiotics not only to DNA but also to proteins such as albumin) are key technologies to this kind of analytical and data interpretation process.
- c) The analysis of the biomarker data (including results on biomarkers of exposure, effects and individual susceptibility) results in the integrated assessment of risk factors. Use of information on risk factors with molecular dosimetry data (i.e. estimation of the actual internal and biologically effective dose of xenobiotic substance found in the target organ and, indeed, perturbing cellular response) enables population risk studies to be done, by converting generic exposure profiles into population risk metrics having taken into account inter-individual variability of response and exposure uncertainty.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	50/61

The connectivity approach was applied on data from Europe-wide campaigns on environmental and biological monitoring of a virtually ubiquitous mixture of volatile organic compounds, i.e. benzene, ethylbenzene, toluene and o-, m- and p-xylene, aldehydes such as formaldehyde and acetaldehyde and a complex mixture of polyaromatic hydrocarbons, which is typical of combustion products (Kotzias et al., 2009). The full array of –omics technologies outlined above were applied to samples of indoor air and dry blood spots and urine of exposed subjects from almost all European Union capitals. Exposure assessment was completed with detailed time activity diaries and questionnaires regarding smoking and dietary habits (especially with regard to alcohol consumption) (Sarigiannis et al., 2011). Gene expression results were processed using clustering algorithms to derive heat maps demonstrating clearly the differences in the biological perturbations caused by parts of the indoor and ambient airborne chemical mixtures in the sampled sites.

Dry blood spot samples from a subset (n=50) of the population which participated in the personal exposure study were analysed with the –omics technologies and state-of-the-art bioinformatics software. The Agilent MicroArray Express was used for analysis of the blood samples and comparison with controls and GeneSpring, the Agilent data analysis system, and in-house bioinformatics analysis software developed on R and Stata was used for –omics data analysis.

The sample analysis followed a hybrid, tiered approach starting from an agnostic search in whole genome mRNA extracted from the biological samples after appropriate sampling and storage at -80 °C. The population data were clustered by group of dominant source of exposure to airborne chemicals including ambient air in cities, typical airborne mixtures in dwellings, schools, kindergartens and other indoor professional environments (public buildings) and samples from people occupationally exposed to fly ash from coal-fired thermal power plants. $p < 0.005$ was used as test of statistical significance revealing the level of gene expression modulation against the controls – in this case only the genes that were up- or down-regulated by more than two-fold compared to the controls were selected. This agnostic search indicated the presence of exposure-specific signatures using gene expression data, especially when considering not just the number but also the loci of the genes that showed the most important differences in expression levels after exposure to the respective mixtures of xenobiotics. Both common parts and clearly distinct regions of the genes with significant modulation in expression were identified when comparing biological samples of subjects exposed to ambient air chemicals against the ones of subjects exposed non-professionally to indoor chemicals and professionally to fly ash (Figure 34). Comparative analysis showed that ambient air chemicals in urban settings across Europe affect a wide spectrum of genes (n=376), most of which are up-regulated (n' = 209); a significant number of genes in the same samples was down-regulated (n''=167). Fly ash had an almost as high an effect on the human genome. Here there were fewer genes that showed modulation in expression (n=214). Unlike ambient urban air chemicals, fly ash from power plants resulted in limited up-regulation (n'=66) and more important down-regulation of gene expression (n''=148). Indoor air chemicals had the smallest effect on gene expression modulation. In this case the total number of genes that modulated their expression levels was contained (n=145). Up-regulated genes were dominant (n'=92) and down-regulated genes were relatively limited (n''=53). A number of genes showed very distinct expression patterns when cross-comparing samples from subjects exposed to different types of airborne chemicals. These could be isolated to serve as exposure signatures at the molecular level. Clearly, such genomic biomarkers of exposure do not imply causal associations with adverse health outcomes. However, they may serve as seed information to the formation of biologically plausible mechanistic hypotheses on adverse outcome pathways (AOPs). These hypotheses help in limiting the investigation space for AOPs without compromising the agnostic nature of the initial analyses.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	51/61

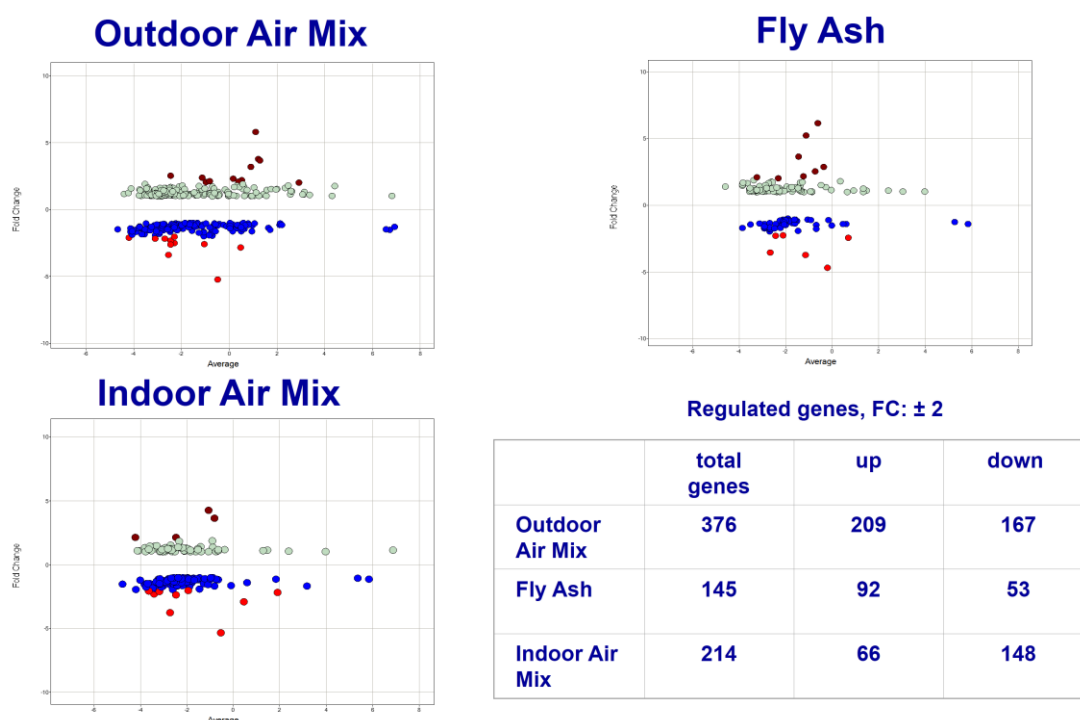



Figure 34: Comparison of gene expression modulation from biosamples of population exposed to different types of airborne chemical mixtures: (a) ambient air chemicals (top left); (b) characteristic airborne chemicals found in indoor settings (top right); (c) chemicals found in fly ash from thermal power plants using fossil fuel (bottom left). The table at the bottom right of the figure shows the total number of gene expressed differentially in subjects exposed to the different airborne chemical mixtures and the number of genes up- and down-modulated respectively

The results of the agnostic transcriptome search on biological samples with regard to the identification of potential genomic signatures that can serve as reliable exposure biomarkers were corroborated by means of in vitro testing. For this reason, samples of the airborne chemicals taken in the various environments considered in this study were extracted and applied on cell lines covering both lung epithelial cells (A549) and skin keratinocytes (Ha-CaT). Cluster analysis of the gene expression results are depicted in the heat map shown in Figure 35. It is very clear that indoor chemicals show distinctly different patterns of gene expression modulation in both types of matrices, especially when compared to the mixture of indoor air chemicals and PAHs from the ambient air in urban settings. It should be noted that the latter corresponds to realistic total exposure patterns of the human population in European metropolitan areas. Exposure to ambient air chemicals only shows intermediate results, indicating that co-exposure to ambient and indoor air chemical mixtures in cities may have more than additive effects on gene expression modulation.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	52/61

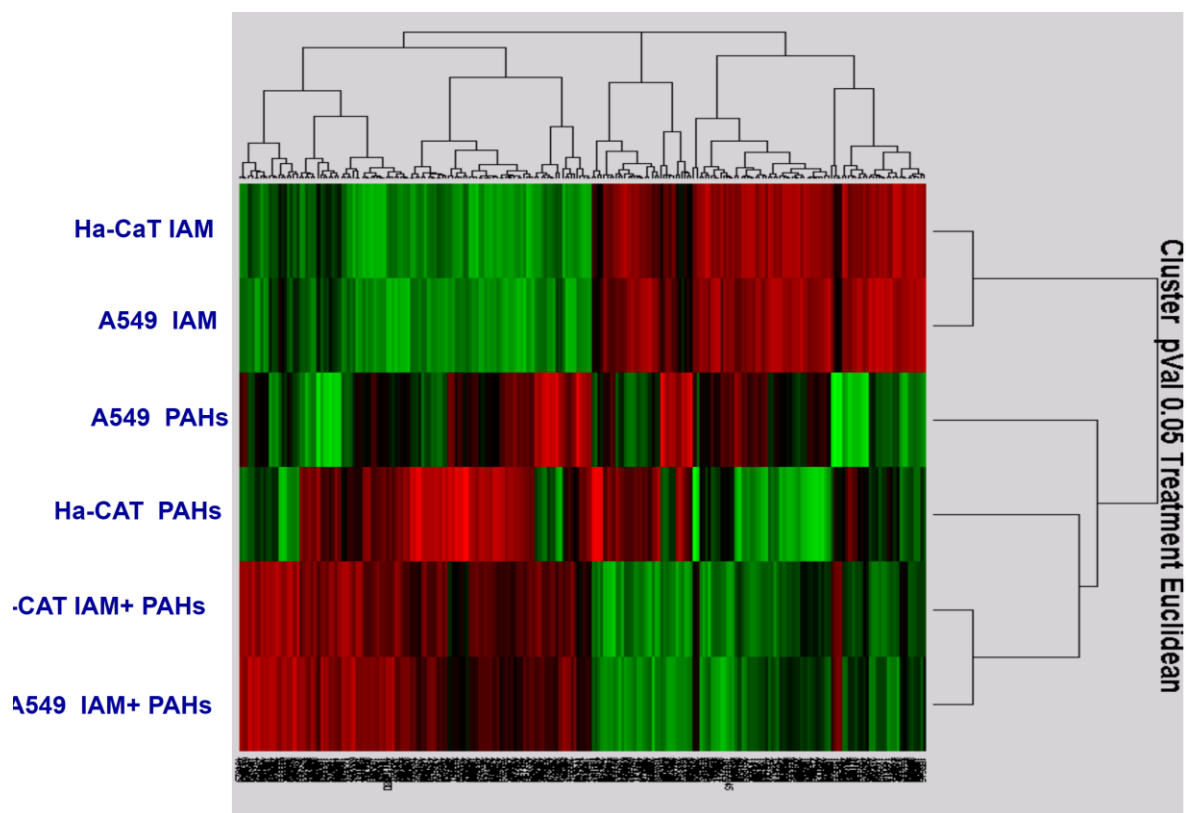



Figure 35: Heat map representing comparative cluster analysis of airborne chemical mixtures on dermal (Ha-CAT) and lung epithelial (A549) cells. IAM: indoor air mixture; PAHs: the mixture of 19 polycyclic aromatic hydrocarbons commonly found in the ambient air in European metropolitan areas

After the agnostic tier it is possible to identify not only single genes that have shown significant modulation in expression levels, but also determine the biological pathways that are regulated by gene networks that were significantly modulated with regard to their induction levels from exposure to xenobiotics. Pathway analysis using Agilent GeneSpring and the on-line PANTHER – Protein ANalysis THrough Evolutionary Relationships) Classification System (<http://www.pantherdb.org>) showed that two key pathways, the p53 (regulating cell cycle, senescence and apoptosis) and oxidative stress were differentially modulated from specific chemical families such as aldehydes, while specific genes or gene sequences could be characterized as molecular markers of exposure. In Figure 36 comparative analysis of the gene regulation induction after exposure to different chemical classes of xenobiotics, which comprise a large part of the indoor air mixture found in European dwellings is shown, focusing on the genetic network regulating oxidative stress response at the cellular level. More specifically, the samples of indoor air chemical mixtures taken from different European cities were decomposed and sequentially extracted to analyze separately the effect on gene expression of major chemical families, including: (a) carbonyl compounds such as aldehydes, (b) phenols and other aromatics, and (c) terpenes.

 FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	53/61

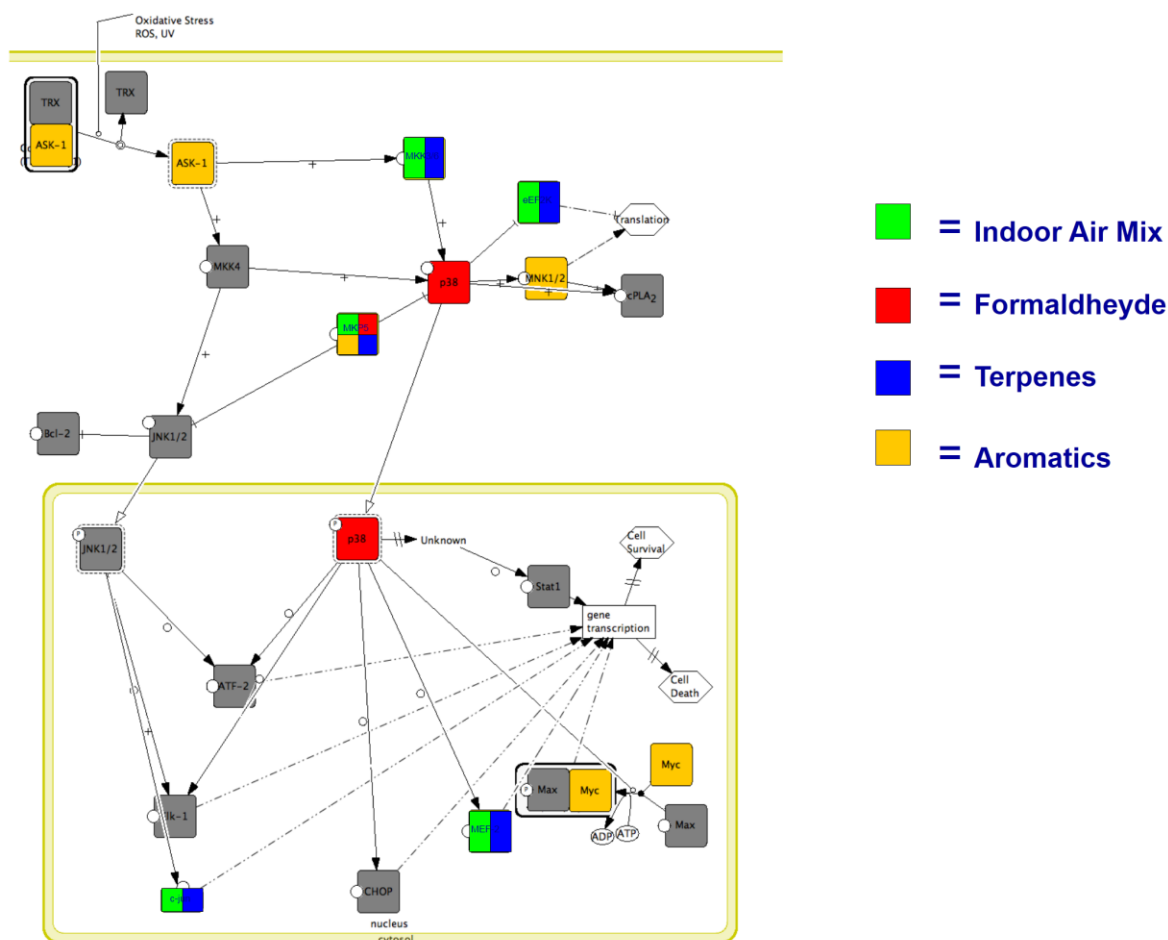



Figure 36: Modulation of the oxidative stress pathway in blood samples of subjects exposed to different airborne chemical families reveals molecular signatures of co-exposure

Aromatics had the highest impact in terms of inducing oxidative stress responses. Terpenes had the second highest impact in terms of activation of parts of the oxidative stress regulatory network. Aldehydes, and in particular formaldehyde, induced among others a specific gene, p38, which is central to the regulation of oxidative stress response in human cells. This gene (p38) could serve as a reliable biomarker of oxidative stress (intermediate effect) associated to exposure to formaldehyde, since no other treatment seemed to modulate its expression levels.

Similar results were found when exploring the modulation of the gene network that regulates the function of the p53 pathway, controlling thus cell cycle and death and thus the onset of carcinogenesis in case of faulty operation and down-regulation. In this case the indoor air chemical mixture as a whole has the highest impact. It seems to induce the expression of the overall regulatory network thus leading to significant up-regulation compared to the controls. Aromatics and terpenes have similar size effects on gene network regulation. Terpenes are more involved in processes regulating apoptosis, as well as inhibition of angiogenesis and metastasis. Aldehydes are involved in processes leading to inhibition of angiogenesis and metastasis too. Thus, different chemical families show different toxicity profiles, not only when associating them with phenotypes of disease such as

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	54/61

cancer, but also, and possibly more importantly, when elucidating their role in biological pathways of toxicity that may (or may not) lead to adverse health outcomes.

From the chemical speciation of the air samples taken in the different European urban settings in this study, typical airborne chemical mixtures were identified for indoor and ambient air as a function of latitude. For example, with regard to volatile organic chemicals, compounds that are essentially ubiquitous in Europe, air mixtures were richer in benzene in the south compared to the north of Europe, where toluene seemed to be more abundant. Following through to the biological processes that were perturbed by exposure to these airborne chemical mixtures we found that during acute (short-term) exposure signal transduction and mRNA transcription were modulated the most following an inverse dose-response function. When chronic (longer-term) exposure results were analyzed, protein metabolism, mRNA transcription regulation and cell proliferation and differentiation were the main mechanisms that were modulated following a normal dose-response behavior. The salient results of this tier of the connectivity analysis are given in Figure 37 for mixture A (the benzene-rich mixture, typical of the European south) and for mixture B (the one that was relatively richer in toluene and other chemically similar solvents, typical of central and northern Europe).

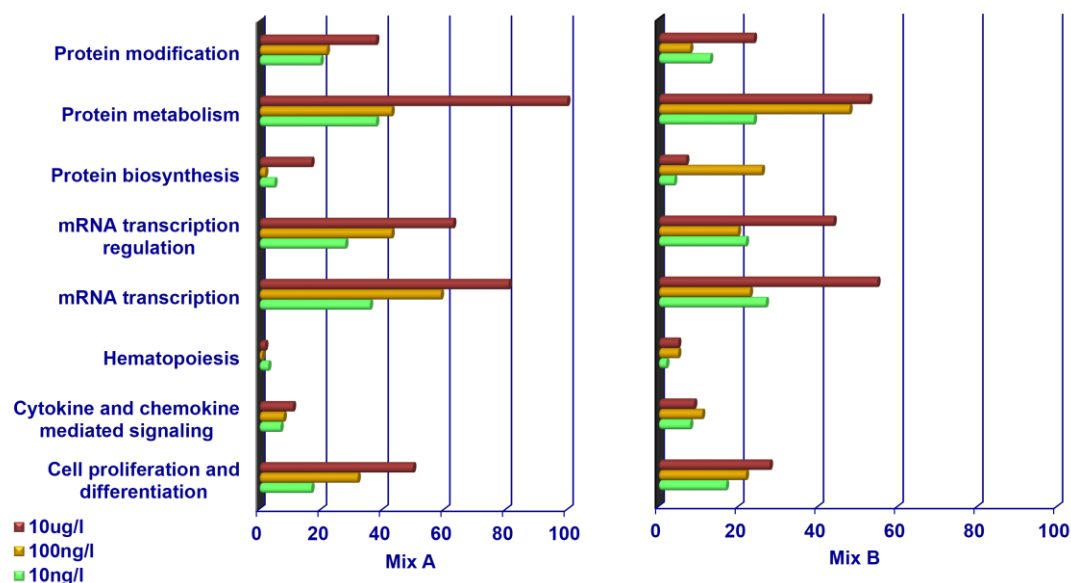



Figure 37: Biological processes induced by the two characteristic types of indoor air mixtures in Europe

In addition, the mixtures that were richer with non-carcinogens were the ones that had the highest impact on the biological processes induced during short-term exposure; on the contrary, mixtures richer with carcinogens such as benzene induced a higher response in terms of biological process induction to the individuals who were exposed longer. This behavior was confirmed by untargeted metabolomics profiling, which showed a relative increase in benzene metabolites such as s-mercapturic acid as well as free, non-metabolized benzene. Phenotypic observations confirm that chronic exposure to carcinogenic VOCs such as benzene could increase the risk of leukemia. Taking into account the metabolic processes and interactions (e.g. competitive inhibition) that regulate the effective metabolism of the VOCs to which the European population was exposed we estimated the actual risk of cancer from the combined exposure to such airborne chemical mixtures released from

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery		Security:
	Author(s): Denis A. Sarigiannis	Version: 1	55/61

fuel and consumer products. Focusing even further on the molecular underpinning of the induced modulation of the human capacity to metabolize VOCs under co-exposure conditions, targeted analysis of the CyP450 was performed using Taqman Fluidigm microarrays. Results showed a significant variation in the expression of genes involved in the coding of the enzymes involved in benzene and toluene metabolic chains. This variation results in reduction of the amount of available enzymes and thus enhances the effect of competitive inhibition for the limited amount of receptor sites at the metabolically active tissues. The results show suppression of CyP450 metabolic capacity after co-exposure to the four VOCs (Figure 38). These findings were translated into quantitative change in the metabolic rates captured in the biology-based dose-response (BBDR) model developed in-house to estimate cancer risk by applying dynamic flux balance analysis and coupling the gene regulatory network with the ADME model.

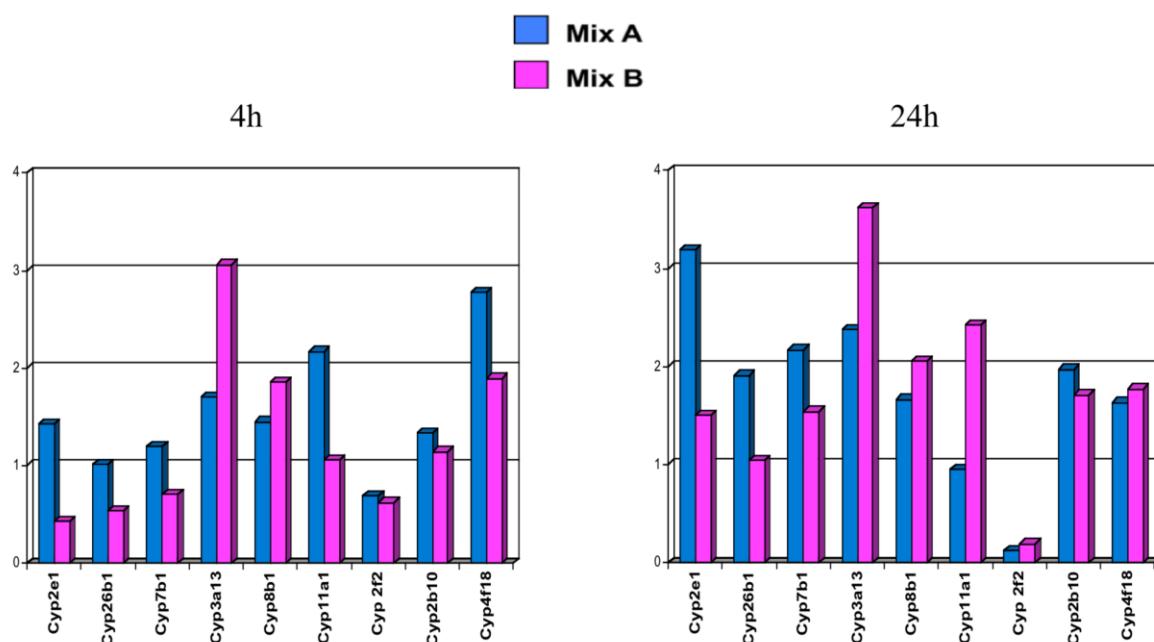




Figure 38: Comparison of CyP450 expression modulation for benzene-rich (black bar) and toluene-rich (white bar) quaternary BTEX mixtures after 4 (left figure) and 24 h (right figure) of exposure

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	56/61

8 References

- Aebersold, R., Mann, M., (2003). Mass spectrometry-based proteomics. *Nature*. 422: 198–207.
- Agrawal, R., Srikant, R.,, Fast Algorithms for Mining Association Rules in Large Databases. VLDB Conference, 1994, pp. 487–499.
- Allison, D. B., Cui, X., Page, G. P., Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics*. 7: 55-65.
- Altman, R. B., Raychaudhuri, S. (2001). Whole-genome expression analysis: Challenges beyond clustering. *Curr. Opin. Struct. Biol.* 11: 340-347.
- Bandyopadhyay, S. (2005). An efficient technique for superfamily classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection. *Fuzzy Sets and Systems*. 152: 5-16.
- Barla, A., Jurman, G., Riccadonna, S., Merler, S., Chierici, M., Furlanello, C. (2008). Machine learning methods for predictive proteomics. *Briefings in Bioinformatics*. 9: 119-128.
- Belacel, N., Čuperlović-Culf, M., Laflamme, M., Ouellette, R.,, (2004). Fuzzy J-Means and VNS methods for clustering genes from microarray data. *Bioinformatics*. 20: 1690–1701.
- Ben-Dor, A., Shamir, R., Yakhini, Z., (1999). Clustering gene expression patterns. *J. Comput. Biol.* 63: 281–297.
- Blankenbecler, R., Ohlsson, M., Peterson, C., Ringner, M.,, Matching protein structures with fuzzy alignments. Proc. Natl. Acad. Sci. U. S. A., Vol. 100, 2003, pp. 11936–11940.
- Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta R.,, 2009. Metalearning: applications to data mining. Springer, Berlin.
- Brazdil, P., Soares, C., de Costa, P., (2003). Ranking learning algorithms: using IBL and meta-learning on accuracy and time results. *Mach Learn*. 50: 251–277.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*. 24: 123-140.
- Breiman, L. (1998). Arcing classifiers (with discussion). *The Annals of Statistics*. 26: 801–849.
- Breiman, L. (2001). Random forests. *Machine Learning*. 45: 5-32.
- Breiman, L., Friedman, J., Olsen, R., Stone, C., , 1984. Classification and Regression Trees. California.
- Carleos, C., Rodriguez, F., Lamelas, H., Baro, J.A., (2003). Simulating complex traits influenced by genes with fuzzy-valued effects in pedigreed populations. *Bioinformatics*. 19: 144–148.
- Chang, B., Halgamuge, S.K. , (2002). Protein motif extraction with neuro-fuzzy optimization. *Bioinformatics*. 18: 1084–1090.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	57/61

Cordón, O., Gomide, F., Herrera, F., Hoffmann, F., Magdalena, L., (2004). Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy Sets and Systems*. 141: 5–31.

Dasgupta, A., Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*. 93: 294-302.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*. 18: 735-746.

Dembele, D., Kastner P., (2003). Fuzzy C-means method for clustering microarray data. *Bioinformatics*. 19: 973–980.

Dong, G., Zhang, X., Wong, L., Li, J., CAEP: Classification by aggregating emerging patterns. In: Springer-Verlag, (Ed.), Proceedings of the Second International Conference on Discovery Science, 1999, pp. 30–42.

Dubes, R., 1988. Algorithms for Clustering Data.

Duh, M. S., Walker, A. M., Ayanian, J. Z. (1998). Epidemiologic interpretation of artificial neural networks. *Am. J. Epidemiol.* 147: 1112-1122.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P, Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 82.

Fraley, C., Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*. 41: 586-588.

Freund, Y., Schapire, R.,, Experiments with a new boosting algorithm. Proceedings of the Thirteenth National Conference on Machine Learning, 1996, pp. 148–156.

Gasch, A., Eisen, M., (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* 3: 1–22.


Han, J., Pei, H., Yin, Y.,, Mining Frequent Patterns without Candidate Generation. Conf. on the Management of Data. ACM Press, Dalas, 2000.

Han, J., Pei, J., Yin, Y., Mao, R., , Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 2003.

Hartuv, E., Shamir, Ron., (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters*. 76: 175–181.

Heger, A., Holm, L., (2003). Sensitive pattern discovery with ‘fuzzy’ alignments of distantly related proteins. *Bioinformatics*. 2003: i130–i137.

Herrero, J., Valencia, A., Dopazo, J., (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*. 17: 126–136.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	58/61

Hirschhorn, J. N., Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*. 6: 95-108.

Huang, Y., Li, Y., (2004). Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*. 20: 21–28.

Jiang, D., Pei, J., Zhang, A., Interactive exploration of coherent patterns in time-series gene expression data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 565-570.

Jiang, D., Pei, J. and Zhang, A. . D, A Density-based Hierarchical Clustering Method for Timeseries

Gene Expression Data. In: BIBE2003, (Ed.), 3rd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, 2003.

Jones, D. T. (2001). Protein structure prediction in genomics. *Briefings in Bioinformatics*. 2: 111-125.

Kalousis, A., Theoharis, T., (1999). NOEMON: design, implementaion and performance results of an intelligent assistant for classifier selection. *Intell Data Anal* 5: 319–337.

Kaufman, L., Rousseeuw, P.J.,, 1990. Finding Groups in Data: an Introduction to Cluster Analysis.

Kitano, H. (2002a). Computational systems biology. *Nature*. 420: 206-210.

Kitano, H. (2002b). Systems biology: A brief overview. *Science*. 295: 1662-1664.

Kohonen, T., 1984. Self-Organization and Associative Memory. Spring-Verlag, Berlin.

Kotzias, D., Geiss, O., Tirendi, S., Josefa, B. M., Reina, V., Gotti, A., Graziella, C. R., Casati, B., Marafante, E., Sarigiannis, D. (2009). Exposure to multiple air contaminants in public buildings, schools and kindergartens-the European indoor air monitoring and exposure assessment (airmex) study. *Fresenius Environmental Bulletin*. 18: 670-681.


Kuncheva, L., 2004. Combining Pattern Classifiers:Methods and Algorithms. Wiley.

Larranaga, P., Calvo, B., Santana. R., Bielza. C., Galdiano, J., (2003). Machine learning in bioinformatics. *Briefings in Bioinformatics*. 7: 86-112.

Li, Y., Chen, L. (2014). Big Biological Data: Challenges and Opportunities. *Genomics Proteomics Bioinformatics*. 12: 187-189.

Liao, K. H., Dobrev, I. D., Dennison Jr, J. E., Andersen, M. E., Reisfeld, B., Reardon, K. F., Campaign, J. A., Wei, W., Klein, M. T., Quann, R. J., Yang, R. S. H. (2002). Application of biologically based computer modeling to simple or complex mixtures. *Environ. Health Perspect*. 110: 957-963.

Lukac. R., P., KN., Smolka, B., Venetsanopoulos, AN., (2005). cDNA microarray image processing using fuzzy vector filtering framework. *Fuzzy Sets and Systems*. 152: 17–35.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	59/61

Martinez, R., Collard, M., (2007). Extracted knowledge: Interpretation in mining biological data, a survey. *International Journal of Computer Science and Applications*. 1: 1-21.

Mason, A. M., Borgert, C. J., Bus, J. S., Moiz Mumtaz, M., Simmons, J. E., Sipes, I. G. (2007). Improving the scientific foundation for mixtures joint toxicity and risk assessment: Contributions from the SOT mixtures project-Introduction. *Toxicol. Appl. Pharmacol.* 223: 99-103.

McQueen, J. B., Some methods for classification and analysis of multivariate observations. In: U. o. C. Press, (Ed.), Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. Univ.of California Press, Berkeley, 1967, pp. 281–297.

Olden, J. D., Joy, M. K., Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol. Modell.* 178: 389-397.

Pearl, J., Verma, T. S., A theory of inferred causation. Principles of Knowledge Representation and Reasoning Second International Conference, 1991, pp. 441–452.

Phan, J. H., Quo, C. F., Wang, M. D., Chapter 4 Functional genomics and proteomics in the clinical neurosciences: data mining and bioinformatics. *Prog. Brain Res.*, Vol. 158, 2006, pp. 83-108.

Pleil, J. D. (2012). Categorizing biomarkers of the human exposome and developing metrics for assessing environmental sustainability. *Journal of toxicology and environmental health. Part B, Critical reviews*. 15: 264-80.

Quinlan, J., 1986. C4.5: Programs for machine learning. San Mateo.

Sarigiannis, D., Marafante, E., Gotti, A., Reale, G. C. (2009). Reflections on new directions for risk assessment of environmental chemical mixtures. *International Journal of Risk Assessment and Management*. 13: 216-241.


Sarigiannis, D. A., Karakitsios, S. P., Gotti, A., Liakos, I. L., Katsoyiannis, A. (2011). Exposure to major volatile organic compounds and carbonyls in European indoor environments and associated health risk. *Environ. Int.* 37: 743-765.

Schapire, R., Freund, Y., Bartlett, P., Lee, WS., (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*. 26: 1651–1686.

Schlosshauer, M., Ohlsson, M., (2002). A novel approach to local reliability of sequence alignments. *Bioinformatics*. 18: 847–854.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y. M., Denys, M., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Gomez, M., Gonzales, E., Haydu, L., Lopez, F., Ramirez, L., Retterer, J., Rodriguez, A., Rogers, S., Salazar, A., Tsai, M., Myers, R. M. (2004). Quality assessment of the human genome sequence. *Nature*. 429: 365-368.

Seno, M., Karypis, G., LPMiner: An Algorithm for Finding Frequent Itemsets Using Length-Decreasing Support Constraint. 1st IEEE Conference on Data Mining, 2001.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	60/61

Shamir, R., Sharan, R., Click: A clustering algorithm for gene expression analysis. In: AAAI Press., (Ed.), 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00), 2000.

Shatkay, H., Edwards, S., W., W., M., B., Genes, themes, microarrays: using information retrieval for large-scale gene analysis. Proc. ISMB 2000, pp. 340–347.

Sherlock, G. (2000). Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* 12: 201–205.

Silva, A., Cortez, P., Santos, M. F., Gomes, L., Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit. *Artif. Intell. Med.* 43: 179-193.

Smith, A. E., Nugent, C. D., McClean, S. I. (2003). Evaluation of inherent performance of intelligent medical decision support systems: Utilising neural networks as an example. *Artif. Intell. Med.* 27: 1-27.

Spirtes, P., Glymour, C., Scheines, R., 1993. Causation, prediction, and search.

Tamayo, P., Solni, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R. I., Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Natl. Acad. Sci.*, Vol. 96, 1999, pp. 2907–2912.

Todorovski, L., Blockeel, H., Dzeroski, S., Ranking with predictive clustering trees. In: M. H. Elomaa T, Toivonen H, (Ed.), Proceedings of the 13th European conference on machine learning. Springer, 2002, pp. 444–455.

Todorovski, L., Džeroski, S., (2003). Combining classifiers with meta decision trees. *Machine Learning.* 50: 223-249.

Tomida, S., Hanai, T., Honda, H., Kobayashi, T. , (2002). Analysis of expression profile using fuzzy adaptive resonance theory. *Bioinformatics.* 18: 1073–1083.

Torres, A., Nieto, JJ.m., (2003). The fuzzy polynucleotide space: basic properties. *Bioinformatics.* 19: 587–592.


Tung, W. L., Quek, C., Cheng, P. (2004). GenSo-EWS: A novel neural-fuzzy based early warning system for predicting bank failures. *Neural Networks.* 17: 567-587.

Valencia, A., Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* 12: 368-373.

Wall, R., Cunningham, P., Walsh, P., Byrne, S. (2003). Explaining the output of ensembles in medical decision support on a case by case basis. *Artif. Intell. Med.* 28: 191-206.

Wang, D., Lee, NK., Dillon, TS., (2003). Extraction and optimization of fuzzy protein sequence classification rules using GRBF neural networks. *Neural Information Processing—Letters and Reviews.* 1: 53–59.

Wang, J. (2008). Computational biology of genome expression and regulation - A review of microarray bioinformatics. *J. Environ. Pathol. Toxicol. Oncol.* 27: 157-179.

 HEALS FP7-ENV-2013-603946	D7.1 - Data infrastructure and data mining		
	WP7: Novel bioinformatics for predictive biomarker discovery	Security:	
	Author(s): Denis A. Sarigiannis	Version: 1	61/61

Webb, G., Zheng, Z., Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. *IEEE Transactions on Knowledge and Data Engineering.*, Vol. 16, 2004, pp. 980–991.

Wild, C. P. (2005). Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention*. 14: 1847-1850.

Wong, J. W. H., Cagney, G., Cartwright, H. M. (2005a). SpecAlign - Processing and alignment of mass spectra datasets. *Bioinformatics*. 21: 2088-2090.

Wong, J. W. H., Durante, C., Cartwright, H. M. (2005b). Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal. Chem.* 77: 5655-5661.

Woolf, P., Wang, Y., (2000). A fuzzy logic approach to analyzing gene expression data. *Physiol Genomics*. 3: 9–15.

Workman, C. T., Mak, H. C., McCuine, S., Tagne, J. B., Agarwal, M., Ozier, O., Begley, T. J., Samson, L. D., Ideker, T. (2006). A systems approach to mapping DNA damage response pathways. *Science*. 312: 1054-1059.

Xing, E. P., Karp, R.M., (2001). Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*. 17: 306–315.

Zadeh, L. A., (1965). Fuzzy Sets. *Information and Control*. 8: 338-353.

Zeng, J., Zhu, S., Yan, H. (2009). Towards accurate human promoter recognition: A review of currently used sequence features and classification methods. *Briefings in Bioinformatics*. 10: 498-508.