Proteomic Based Identification of Cancer Biomarkers: The LOCCANDIA Integrated Platform

M. Kalaitzakis[1], V. Kritsotakis[1], P. Grangeat[2], C. Paulus[2], L. Gerfault[2] , M. Perez[3], C. Reina[3], G. Potamias[1], M. Tsiknakis[1], D. Kafetzopoulos[4], Pierre-Alain Binz[5]

Abstract—**Pancreatic cancer is the fourth leading cause of cancer death in the United States. Consequently, identification of clinically relevant biomarkers for the early detection of this cancer type is urgently needed. In recent years, proteomics profiling techniques combined with various data analysis methods have been successfully used to gain critical insights into processes and mechanisms underlying pathologic conditions, particularly as they relate to cancer.**

The LOCCANDIA (Lab-On-Chip based protein profiling for CANcer DIAgnosis) project is primarily concerned with validating the application of plasma protein profiling for early pancreatic cancer diagnosis by means of developing an innovative nano-technology based (lab-on-a-chip) platform integrated in a full proteomics analysis chain.

This paper describes the integrated clinico-proteomic information management and analysis platform. In particular it focuses on discussing the underlying methodologies and technological aspects of key SW modules, i.e. the data preprocessing and profile reconstruction as well as the classification modules.

I. Introduction

The human plasma proteome holds the promise of a revolution in disease diagnosis and therapeutic monitoring. The plasma protein analysis aims to characterize the proteomic status of cells and in particular to define the degree of their disorder according to their expression level pattern. This is in particular highly relevant to the effort that has been done in associating specific protein marker levels in patients' blood with the different cancer stages. One major breakthrough comes from the utilization of multi-protein disease markers instead of single protein analytes and the detection of all the isoforms of the selected proteins.

Gastric cancers, such as pancreatic cancers, are among the most frequently observed severe diseases in developed countries [1]. These types of cancers are detected by expensive diagnostic imaging methods at a late stage resulting in poor prognosis and high mortality rate since the only effective therapy is an early resection of the tumors.

Since the clinical manifestations of pancreatic cancer, except obstructive jaundice, are often not apparent until the advanced stages of the disease, and the anatomical location of the pancreas deep in the abdomen makes physical and ultrasonic detection of pancreatic cancer difficult, about 95% of all cases are diagnosed in stage III or IV, and the 5-year survival rate of pancreatic cancer patients is the lowest among patients with common solid tumors.

Human blood serum and plasma contain a large variety of proteins, and their relative abundance and modification may precisely reflect the disease status of organs and tissues. Recent advances in MS-based proteomic technologies coupled with bioinformatics may revolutionize medical diagnosis and cancer screening. Mass spectrometry approaches [2] are very attractive to detect protein panels and protein isoforms in a sensitive way. However, the application to clinical diagnosis is still at its beginning [3]. The need for new and relatively simple devices to allow for the translation of these research results to clinical practice is urgent.

The LOCCANDIA (Lab-On-Chip based protein profiling for CANcer DIAgnosis) project is primarily concerned with validating the application of plasma protein profiling for early pancreatic cancer diagnosis by means of developing an innovative nano-technology based (lab-on-a-chip) platform integrated in a full proteomics analysis chain [4].

This paper describes the integrated clinico-proteomic information management and analysis platform. In particular it focuses on discussing the underlying methodologies and technological aspects of the integrated modules, i.e. the data preprocessing and profile reconstruction and the classification modules.

II. Information Modeling

The LOCCANDIA project integrates a full proteomic analysis chain, from blood sample to diagnostic information, combining bio, nano and information related technologies (fig. 1).

The BIO part of the analysis chain gets as input blood sample, and provides the NANO part with a selected protein mixture that will be further analyzed. NANO is the part related to Lab-on-Chip and Mass Spectrometry. This part is divided in several modules; target protein mixture treatment, digestion, liquid chromatography, electrospray ionization, and mass spectrometry. The INFO part is related to the

[1]Institute of Computer Science, Foundation for Research and Technology-Hellas (ICS-FORTH), Heraklion, Crete, Greece
{tsiknaki, mkalaitz, vkrits, potamias}@ics.forth.gr
[2]CEA-LETI MINATEC, Grenoble, France
{pierre.grangeat, caroline.paulus, laurent.gerfault }@cea.fr
[3]ATOS Research and Innovation, Biotechnology and Health Unit, Madrid, Spain {manuel.perez, carmen.reina}@atosresearch.eu
[4]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas (ICS-FORTH), Heraklion, Crete, Greece
{kafetzo@imbb.forth.gr}
[5]Geneva Bioinformatics (GeneBio) SA, Geneva, Switzerland
{pierre-alain.binz@genebio.com}

supporting information technology infrastructure and is utilized through the LOCCANDIA Information Management System (LIMS). LIMS is responsible for documenting real life events and processing primary information assets for knowledge discovery.

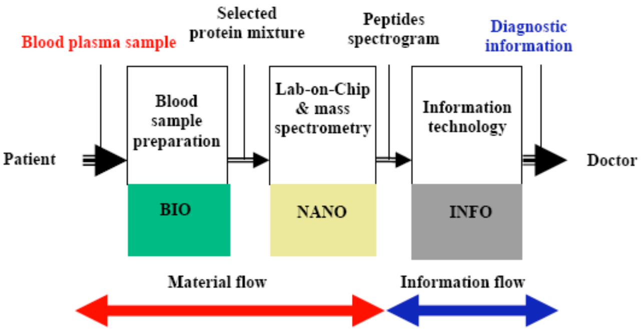


Fig. 1. The LOCCANDIA diagnostic analysis chain

Past experience has shown that multivariate analyses (as proteomics generally is) require an enormous wealth of data. This requires the design and implementation of a "proteomics database", which must adhere to strict rules ensuring the desired clarity and quality. Therefore such a database was designed and developed, by building on the outputs of existing standardization initiatives.

One of the fundamental requirements, arising from the discussion above, is that of integration of information across multiple information domains, one of genomics/proteomics research and the other of clinical practice. Building on such a requirements base, a proteomic information management workflow involves collecting, indexing, searching, and analyzing data generated from a wide variety of instrument, robotics, and software platforms. In order to automate this process, the solution must support third-party integration, management, and reporting in a consistent, logical manner.

LIMS has been designed with the following high-level goals in mind; scalability, availability, performance, security, extensibility and modularity.

A major challenge in the systematic capture of protein expression data is the diversity of experimental technologies and data formats in the field. Various XML (eXtensible Markup Language) standards for proteomics have been developed in order to facilitate the capture, analysis, and distribution of proteomics data. mzXML [5], [6] is a XML based common file format for proteomics mass spectrometric data. The intent of mzXML is to encapsulate unprocessed, raw peak lists. Therefore, mzXML is the data format used as a data input on the profile reconstruction module.

III. LIMS PLATFORM

The LOCCANDIA Information Management System (LIMS) is a web-based application, responsible for the storage, examination and manipulation of clinical and proteomic data. LIMS acts as a mediation platform for the integration and data exchange between several data analysis tools. Its ultimate objective is to intelligently correlate clinical and proteomic information towards LOCCANDIA's goal of early pancreatic cancer diagnosis [7].

During the project's initial phases it was decided to employ a user-driven iterative and incremental development (IID) [8]. Each iteration should be a self-contained mini-project composed of activities such as requirements analysis, design, programming, and test. Likewise, test-driven development techniques [9] were adopted in order to achieve the desired improvement and functionality of the system.

LIMS was designed and developed using the "three-tier" client-server software architecture. The application uses an authorization security mechanism for user authentication. The web-based graphical user interface consists of a large number of view tables and input forms, used for data view and manipulation (fig.2).

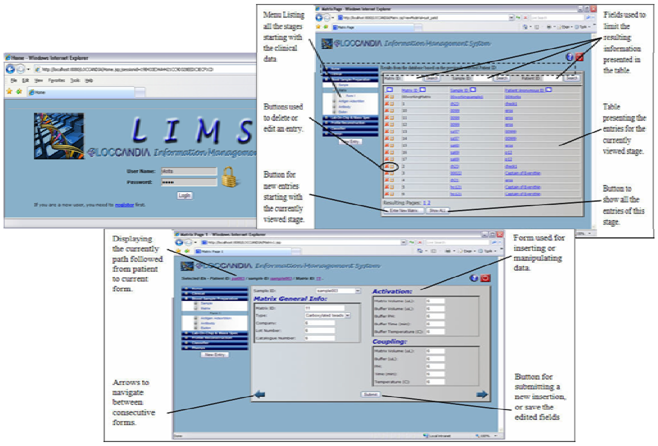

L I M S
LOCCANDIA Information Management System
User Name:
Password:
Login
Menu Listing all the stages starting with the clinical data
Fields used to limit the resulting information presented in the table
Buttons used to delete or edit an entry.
Table presenting the entries for the currently viewed stage.
Button for new entries starting with the currently viewed stage.
Button to show all the entries of this stage.
Displaying the currently path followed from patient to current form.
Form used for inserting or manipulating data.
Matrix General Info:
Activation:
Coupling:
Submit
Arrows to navigate between consecutive forms.
Button for submitting a new insertion, or save the edited fields

Fig. 2 LIMS clinic-proteomic platform

IV. DATA PREPROCESSING AND PROFILE RECONSTRUCTION

A. Module Description

This module estimates the concentration of the unknown proteins [10]. Its functionalities can be divided in two major operations. Data pre-processing, which is responsible for correcting the time shifts between the experiments and profile reconstruction that is applied in order to quantify the proteins.

The objective of the proteomic profile reconstruction task is to develop the methodology and the signal reconstruction software to recover quantitative molecular concentration profiles from the raw measurements, and to evaluate the results on both synthetic and experimental data.

To compensate for systematic differences due to sample loadings and instrument errors, raw proteomics data have to be pre-processed before any feature selection method and classification algorithm can be applied. Three major pre-processing procedures were applied to our dataset: baseline adjustment, normalization and kernel smoothing (see fig 3).

Pre-processing is initially applied, as it standardizes data and corrects the time delay in retention time among spectrograms [11]. To achieve that a block matching algorithm has been implemented. Contrary to classical correction algorithms, block matching does not use Dynamic

Time Warping and thus is not restricted to 1D signal. In fact the proposed method adopts the usual translational motion models and is applied on Liquid Chromatography-Mass Spectrometry "images".

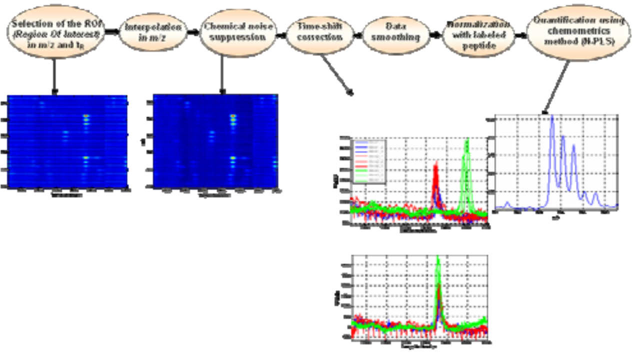

Selection of the ROI (Region Of Interest) in m/z and t_R
Interpolation in m/z
Chemical noise suppression
Time-shift correction
Data smoothing
Normalization with labeled peptide
Quantification using chemometrics method (N-PLS)

Fig. 3. Computational chain for the profile pre-processing & reconstruction

The Profile reconstruction SW module is developed iteratively in two versions. The first version is based on the inverse problem approach and relies on developing a functional model that associates the concentration profile of the unknown molecular with the measurements of the spectrograms. Unlike other commonly used approaches, it is based on chemometric methods [12] and thus has the advantage of increasing the sensitivity and robustness to noise and perturbations as it takes into account the whole signal. The second version will introduce physical and chemical parameterization of the system.

Currently, version 1 (V1) of the reconstruction module has been automated and incorporated into the LIMS, while version 2 (V2) is under development at CEA LETI. The following section provides detailed information on the implementation of version 1.

B. Module Input/Output

The profile reconstruction module and pre-processing is a MATLAB script that runs in the MATLAB environment. The MATLAB code includes the following steps:

- Pre-processing (common for V1, V2):
  - Conversion of mzXML
  - Interpolation (in mass over charge for QTOF and retention time for LIT)
  - Time shift correction between the various experiments
  - Data smoothing (to remove some noise)
- Profile reconstruction (V1) with various chemometrics methods.

The LIMS interface only needs to call one MATLAB function "recons_V1". The entire steps have been automated.

The function input arguments are:

- Type of mass spectrometer ('LIT' or 'QTOF').
- Targeted protein concentration in the calibration mixtures.
- Mass over charge ratio of targeted peptides (+/- delta).
- Retention time of targeted peptides (+/- delta).
- Over load in mass of tagged peptide.
- Tagged protein concentration in all the mixtures.
- Experimental files (calibration files followed by prediction files).

The function output arguments are:

- Estimated concentrations of the targeted proteins in the prediction files with the 4 different methods of V1 (CLS, PLS, N-PLS, PARAFAC).

C. Version 1: Non-parametric and linear (chemometrics approach)

The chemometrics approach is used for V1 of the reconstruction algorithm. In literature, chemometrics techniques correspond to black box model based approach. The aim is to extract the relevant information of physicochemical measured data based on the construction and the exploitation of a multivariable model.

Chemometrics techniques model variations of a given number of variables, called *Xvariables* for which the estimate is delicate (in our case, proteins concentration), in function of others variables, called *Yvariables*, easily measurable (2D spectrograms).

Two operations are necessary:

- **Calibration step** realizes the computation of the model *C*. This procedure is done off-line. At this step, the *Xvariables* and the *Yvariables* are known.
- **Prediction step** enables to compute the *Xvariables* using the model estimated at the calibration step and the *Yvariables*.

The implemented function uses the N-way toolbox [13] for MATLAB, an advanced freeware toolbox for fitting multi-way model.

The following methods have been implemented:

- *CLS (Classical Least Square)*

The Classical least Square (CLS) method extends the application of ordinary least square as applied to a single independent variable. CLS method realizes the inversion of two matrices of low dimension.

- *PCR (Principal Component Regression)*

With analysis involving large numbers of independent variables, correlation often exist between the variables. Co-linearity adds redundancy to the regression model since more variables than necessary are used in the model. PCR overcomes this problem by: (a) selecting the smallest number of variables necessary, (b) using the maximum of information contained in the independent variables and (c) choosing independent variables that are not highly correlated with each other.

- *PLS (Partial Least Square)*

The method differs from PCR by including the concentration variable in the data compression and decomposition operations (both X and Y are actively used in the data analysis). For this reason, PLS solution should have better predictive power than PCR when the calibration step does not involve all the components of the system. This serves to minimize the potential effects of Y variables having large variance but which are irrelevant to the calibration model.

– *N-PLS (Multiway Partial Least Square)*

N-PLS is a multiway regression method which realizes PLS on 2D measurements (spectrograms of 2 dimensions) instead of unfolding the matrix into a long-vector. Compared to unfolding methods, the multilinear models are much simpler, because they need fewer parameters and are preferable with regard to simplicity and stabilization of the decomposition. This stabilization potentially gives increased interpretability and better predictions. The algorithm is fast compared to PARAFAC, because it consists of solving eigenvalue problems.

– *PARAFAC (PARallel FACtor analysis)*

PARAFAC is a multiway decomposition method used in chemometrics. An advantage of using multiway methods instead of using unfolding methods is that the estimated models are simple, more robust and easier to interpret. PARAFAC decomposes the array into sets of scores and loading that describes the data in a more condensed form than the original data array. The reason for using multiway methods is not to obtain better fit but rather more adequate, robust and interpretable models.

D. Module Integration

Version 1 of profile reconstruction and pre-processing module is a MATLAB script that runs in the MATLAB environment. JMatLink is a Java library that allows a Java application or servlet to connect to MATLAB by using native methods. JMatLink uses a multi-threading approach to improve performance and handle multiple MATLAB sessions at a time. LIMS uses Jmatlink to call several MATLAB scripts, accept their output and store it into the database.

The user can access the reconstruction module via the LIMS navigation menu. In order to invoke the reconstruction module a certain workflow should be followed (see fig. 4). Initially the user should use the provided upload mechanism to upload all the required mzXML files on the LIMS server. Following that, specific protocol and regions of interest (ROI) should be defined. The Protocol form is used for defining the different mass spectrometer devices used. Protein ROI form is concerned with the Regions of Interest (ROI) per protocol. The next step includes the assignment of the uploaded mzXML files to either calibration or prediction database entities. The calibration entity contains information regarding the known concentrations of the targeted proteins. On the other hand, prediction entities are used to associate the uploaded prediction mzXML files to specific samples stored in the clinico-proteomic part of the database. When the user clicks on the “Run Profile Reconstruction” link, the reconstruction MATLAB function is initiated.

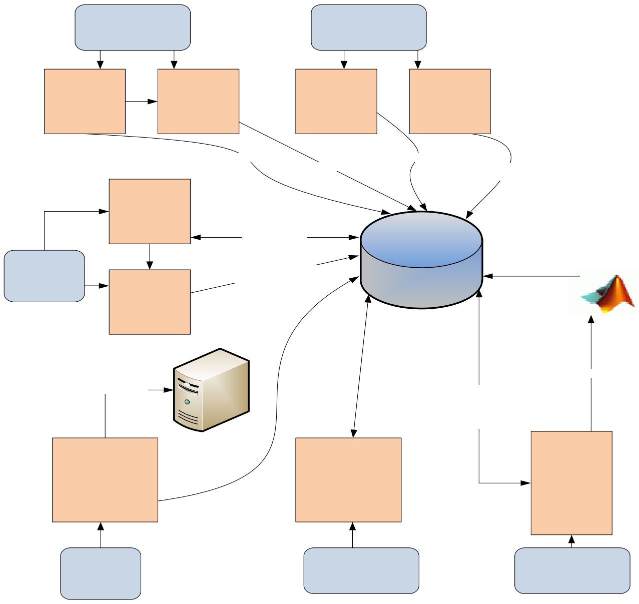
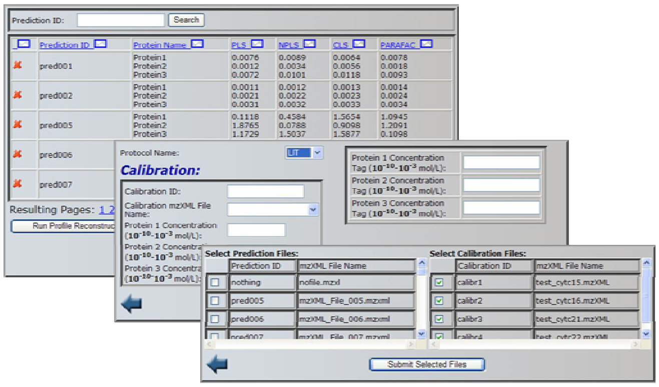


Fig. 4. Process flow and interface diagram

The user then selects the desired prediction and calibration files from a list of available files to be used. By submitting the selected files an algorithm is used to create the appropriate input string for the MATLAB script. The script is then executed and the predicted concentration values per method (CLS, PLS, NPLS, PARAFAC) per prediction file are estimated and stored in the LIMS Database. Part of the integrated profile reconstruction module functionality is shown on fig. 5.



Prediction ID: [ ]

	Prediction ID	Protein Name	PLS	NPLS	CLS	PARAFAC
✗	pred001	Protein1 Protein2 Protein3	0.0076 0.0012 0.0072	0.0089 0.0034 0.0101	0.0064 0.0056 0.0118	0.0078 0.0018 0.0093
✗	pred002	Protein1 Protein2 Protein3	0.0011 0.0021 0.0031	0.0012 0.0022 0.0032	0.0013 0.0023 0.0033	0.0014 0.0024 0.0034
✗	pred005	Protein1 Protein2 Protein3	0.1118 1.8765 1.1729	0.4584 0.0788 1.5037	1.5654 0.9098 1.5877	1.0945 1.2091 0.1098
✗	pred006					
✗	pred007					

Resulting Pages: 1 2

Protocol Name: UT

Calibration:

Calibration ID:
Calibration mzXML File Name:
Protein 1 Concentration (10^{-10}-10^{-3} mol/L):
Protein 2 Concentra (10^{-10}-10^{-3} mol/L):
Protein 3 Concentra (10^{-10}-10^{-3} mol/L):

Protein 1 Concentration Tag (10^{-10}-10^{-3} mol/L):
Protein 2 Concentration Tag (10^{-10}-10^{-3} mol/L):
Protein 3 Concentration Tag (10^{-10}-10^{-3} mol/L):

Select Prediction Files:

	Prediction ID	mzXML File Name
□	nothing	nofile.mzxl
□	pred005	mzXML_File_005.mzxml
□	pred006	mzXML_File_006.mzxml
□	pred007	mzXML_File_007.mzxml

Select Calibration Files:

	Calibration ID	mzXML File Name
☑	calibr1	test_cytc15.mzXML
☑	calibr2	test_cytc16.mzXML
☑	calibr3	test_cytc21.mzXML
☑	calibr4	test_cytc22.mzXML

Fig. 5. LIMS forms and interfaces

V. Classification Module

A. Specific Purpose

At this point of the analytical chain, the time shifts between the experiments have been corrected, the signal has been reconstructed and the quantitative molecular concentration profiles have been recovered. The next step is to use those concentrations in order to characterize the health status of the patients with unknown health conditions.

In achieving this, a Classification Module is used to predict patient outcomes discerning healthy individuals from pancreatic cancer patients by differentiating a given sample between two classes (pancreatic cancer and no pancreatic cancer). In achieving this, it applies specific algorithms to a set of selected proteomic and clinical data, based on modern statistical learning and logistic regression. In addition, Receiver Operating Curve (ROC) [14] analysis is performed for measuring the performance and the accuracy of the different classification methods used.

B. *Machine Learning Classification Algorithms*

Many investigators have applied proteomics technology and data mining methods to identify serum proteomic patterns that can distinguish normal from cancer samples [15].

One of the major challenges for proteomic profiling is the analysis and mining of biologically useful information from the enormous dataset. Due to the high dimensionality of proteomics dataset and their often small sample sizes, non-classical statistical methods for data analysis need to be employed. Therefore, various machine learning classification algorithms have been applied to proteomics data analysis. These include the use of decision tree [16], Bayesian neural network [17], self-organizing map [18], support vector machine [19], linear and quadratic discriminant analysis [20].

The analysis and mining of proteomic information from large dataset had always been challenging. There are several machine learning classification algorithms that have been applied to proteomic data analysis and in some cases they had been limited in terms of efficient procedure and robustness in handling the variance inherent in the proteomic data [15].

C. *Classification Methods*

The classification module developed for the LOCCANDIA information system builds on a variety of classification techniques. The first family of techniques utilizes Support Vector Machines (SVM) [21],[22], which is a set of related supervised learning methods capable of minimizing the empirical classification error and maximizing the geometric margin, whereas the second one makes use of Logistic Regression model, that filters data to a logistic curve by using a number of predictor variables. Several methods based on SVM have been implemented, providing eight different kernels. These are: Linear Kernel, Gaussian Radial Basis Function Kernel, Laplace Radial Basis Function (RBF), Bessel functions, Polynomial Kernel, Hyperbolic tangent kernel, ANOVA radial kernel and Linear Splines.

These eight methods together with a Logistic Regression based one constitute nine different classification approaches.

D. *Support Vector Machine*

SVM uses an implicit mapping Φ of the input data into a high-dimensional feature space defined by a kernel function.

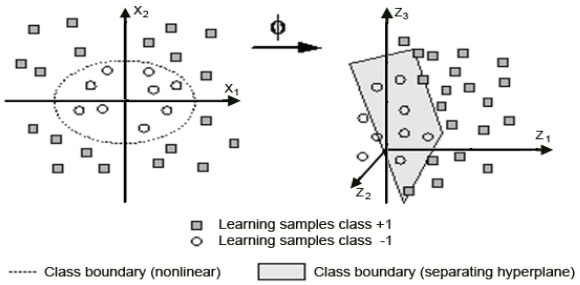

x₂
x₁
ϕ
z₃
z₁
z₂
Learning samples class +1
Learning samples class -1
Class boundary (nonlinear)
Class boundary (separating hyperplane)

Fig. 6. Input space in two dimensions on the left and feature space in three dimensions on the right [23].

The learning then takes place in the feature space, and the data points appear only inside dot products with other points. More precisely, for a projection Φ: X –> H, the dot product <Φ(x), Φ(x')> can be represented by a kernel function k:

$$k(x, x') = <\Phi(x), \Phi(x')>$$

which is computationally simpler than explicitly projecting x and x' into the feature space *H*.
According to the decision function

$$f(x) = \text{sign}(<w, \Phi(x)> + b),$$

a hyperplane <w, Φ(x)> + b = 0 is used in classification, to separate the different classes of data.
The optimal hyper-plane with the optimal classification performance is the one with the maximal margin of separation between the two classes.

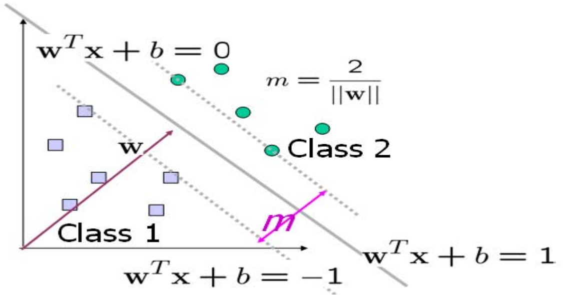

$\mathbf{w}^T\mathbf{x} + b = 0$
$m = \frac{2}{||\mathbf{w}||}$
w
Class 2
Class 1
m
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = -1$

Fig. 7. Margin of separation between the two classes [24].

Similar to most kernel methods, the SVM solution w can be shown in terms of a subset of training patterns that lie on the margin

$$w = \sum_{j=1}^{m} \alpha_i y_i \Phi\ (x_i)$$

The training patterns, called support vectors, are sufficient for capturing the classification problem. Initially the classifier is trained with binary-labeled samples and as soon as training has finished, it can classify new samples. SVM algorithm uses an n-dimensional space to define the hyper-plane that best divides two groups of samples.

E. *Logistic Regression*

Logistic Regression uses several predictor variables that are either numerical or categories. The output of a linear regression can be transformed by using a logit link function

logit p = log *o* = log

$$\frac{p}{1-p} = \beta_0+ \beta_1 x_1+ \beta_2 x_2+\ldots\ldots+\beta_k x_k$$

in order to be more suitable for probabilities.

$$\text{logit}\ (\pi) = z,\ \text{then}\ \pi\ = \frac{e^z}{1+e^z}$$

The inverse of the logit function comprises the logistic function, which maps any value of the z to a proportion value between 0 and 1. Knowing that a certain fact of a data point is true can cause constant change in the odds of the outcome.

F. Kernel Hebbian Algorithm for Iterative Kernel Principal Component Analysis (KPCA)

KPCA [25] is the technique of extracting non-linear structure from data. The input data is initially mapped into a Reproducing Kernel Hilbert Space (RKHS) and then PCA is applied in the space. The direct computation of PCA is achieved using kernel methods and formulating it as the equivalent kernel eigenvalue problem. The principal components are computed by KPCA in a high-dimensional RKHS F. The input space is related to F by a nonlinear map $\Phi : R^n \rightarrow F$.

The inner product of two points mapped by Φ are evaluated by using kernel functions k(x, x') = <Φ(x), Φ(x')>. Considering the fact that PCA is able to be formulated in terms of inner products, it can also be computed in a RKHS.

G. Advantages of the classifier's methodology

The widespread use of SVM in pattern profile classification is not coincidental. SVM provides a simple implementation method for binary classification, it offers modeling freedom through kernel function choice and it is based on machine learning which favors the production of robust classifiers mainly in case of binary data.

However, training is significant for the accuracy of an SVM model. Thus a small training set leads to inefficient classifiers and large homogeneous training set can cause over-fit problems [15]. On the other hand logistic regression is capable of using many predictor variables that can either be numerical or categories.

The classifiers developed have been tested on simulated data sets, including both clinical and proteomic features. Testing the implemented classifiers on real datasets is a key challenge in the work lying ahead of us in the LOCCANDIA project, with the objective of validating the performance of the classification methods selected in the context of the high dimensionality of proteomics data combined with their relatively small sample sizes, which poses a significant challenge to current data mining methodologies.

H. Integration with LIMS

The classification algorithms have been implemented in R environment (http://www.r-project.org) and are embedded into Java (http://java.sun.com/) for the easier integration with the LIMS. The communication between R and Java is achieved with the use of Rserve (http://www.rosuda.org/Rserve/), which allows other software to use facilities of R from various languages.

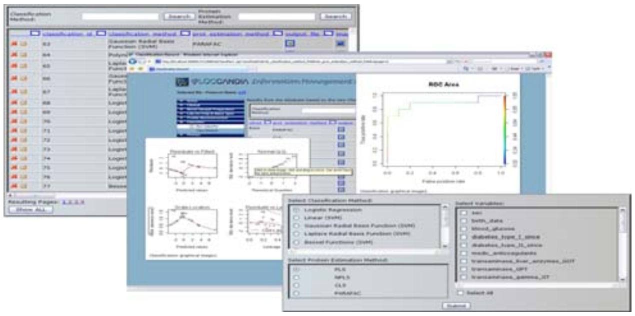


Fig. 8. LIMS interface regarding the classification module.

LIMS provides the required interface for the parameterization and execution of the classification module. The parameters set include the selection of the desired classification method, estimation method and the variables to take place in the classification process. As soon as all the parameters have been set, the LIMS passes the values of patient's selected properties, the protein concentrations and the health status of the patients (where known) to the classifier. Patients with known health status are used as a training set according to the chosen classification method.

There is always a trade-off between sensitivity and specificity because of the different threshold values used in binary prediction. Thus, ROC (Receiver Operating Characteristic) curve is used to plot true/false positive rates or sensitivity/1-specificity for different thresholds. The area under the ROC curve (AUG) equals the probability of correctly classified one pair of samples, each one from a separate class. It has been used as an important measurement of classifier performance. A classifier is considered a preferred classifier compared to the other classifier if it has a larger AUG value. A random classifier has an area of approximately 0.5 under the ROC graph, whereas a perfect classifier has an area of 1.

The LOCCANDIA classification module produces two images and an excel file that holds the predicted health status of a patient (see fig 8). The first image shows the ROC area, whereas the second image presents logistic regression distribution graphs. All outputs are persistently managed in the database along with the record of the classification runs.

It is in our plans to include in future LIMS releases additional data-mining algorithms and methods. For example: feature-selection - for the identification of the most critical clinical and proteomic diagnostic variables, and association rules mining (ARM) - for the discovery of interesting (i.e., with diagnostic value) associations between clinical and proteomic features.

VI. Future Work

A. Protein/peptide identification module

The objective of this module is to identify those proteins and peptides which are not in the initial targeted panel but might be present in the measurement. The implementation of the module is based on the Phenyx (http://www.phenyx-ms.com/) software platform.

Phenyx is a software platform for the identification and characterization of proteins and peptides from mass spectrometry data. It was developed by GeneBio in collaboration with the Swiss Institute of Bioinformatics (SIB) and incorporating the true probabilistic and flexible scoring system OLAV. Phenyx is specifically designed to meet the concurrent demands of high-throughput MS data analysis and dynamic results assessment.

The user will be able to access the Phenyx platform through the LIMS interface. A converter is been developed to be used for processing large mzXML files and identify and only keep the peaklists. The resulting file will then be submitted to Phenyx. Hence, users will be able to use Phenyx in order to validate the existing peptide regions of interest or to identify new ones.

VII. Conclusion

Innovation in LOCCANDIA is based on the seamless integration of a bio, nano and information processing stages for the development of a novel integrated diagnostic system. The information technology part of the project addresses the complexities of the analyzed mixture and is focusing on delivering methods and tools for improving the measurement reliability, and ultimately providing a robust and easy to use system.

The ability to gather and analyse large amounts of data, requires that those data are fully annotated as to their method of generation; including provision of relevant contextual links (e.g., to related data sets); and descriptions of the origin of the biological material under study, the preparative and analytical techniques deployed and data analyses performed. Such annotation is usually referred to as the "metadata" (in essence, this is data about the data). If that annotation is incomplete or ambiguous, any comparison or summary statistics generated may be inaccurate, or even downright misleading; incomplete metadata also precludes the independent assessment of the quality of a method or a data set. To address this issue we provide reporting information whenever data are generated by a particular technique.

As a result, an innovative integrated Clinico-Proteomics computational environment has been developed, combining standard-based informatics systems and best-of-breed computational modules, to support the LOCCANDIA integrated lab-on-chip based diagnostic device.

References

[1] K. Honda, et al, (2005) "Possible Detection of Pancreatic Cancer by Plasma Protein Profiling", Cancer Res; 65:(22). 15 November, pp. 10613-10622.

[2] R. Aebersold, et. al. (2003) "Mass spectrometry-based proteomics", Nature, 422, 2003, pp. 198–207.

[3] H. Mischak, et. al., (2007) "Clinical Proteomics: a need to define the field and to begin to set adequate standards", PROTEOMICS - Clinical Applications, 1: 148-156.

[4] B. Jordan, et. al.,(2008) " LOCCANDIA: Lab-on-Chip Based Protein Profiling for Cancer Diagnosis", in proc. of the 5th International Workshop on Wearable, Micro and Nanosystems for personalized Health (pHealth), 21-23 May, Valencia, Spain.

[5] PG. Pedrioli, et. al. (2004). "A common open representation of mass spectrometry data and its application to proteomics research". Nat. Biotechnol. 22 (11): 1459–66.

[6] SM. Lin, et. al. (2005). "What is mzXML good for?" Expert review of proteomics 2 (6): 839–45.

[7] M. Kalaitzakis, et. al.,(2008) "An Integrated Clinico-Proteomics Information Management and Analysis Platform", in proc. of the 21st IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS 2008). 17-19 June, Jyvaskyla, Finland.

[8] Craig Larman, Victor R. Basili, (2003) "Iterative and Incremental Development: A Brief History," *Computer*, vol. 36, no. 6, pp. 47-56, Jun.

[9] Kent Beck, (2002) "Test Driven Development: By Example", Addison-Wesley Longman, ISBN 0321146530, ISBN-13 978-0321146533.

[10] C. Paulus, et al., (2007) "Chromatographic alignment combined with chemometrics profile reconstruction approaches applied to LC-MS data", in proc. of IEEE EMBC, 23-26 Augoust, Lyon, France.

[11] M. Hilario, et al., (2006) "Processing and classification of protein mass spectra", Mass Spectrom Rev. 25(3): p. 409--449.

[12] S. Wold, (1995) "Chemometrics; what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems", 30: p. 109-115.

[13] C.A. Andersson, and R. Bro, *The N-way Toolbox for MATLAB.* Chemometrics and Intelligent Laboratory Systems. 52(1): p. 1—4.

[14] N.A. Obuchowski (2003). "Receiver operating characteristic curves and their use in radiology". Radiology 229 (1): 3–8. doi:10.1148/radiol.2291010898. PMID 14519861.

[15] G. GE, G. W. Wong, (2008) "Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles", *BMC Bioinformatics* **9:**275

[16] A. Vlahou, J.O. Schorge, B.W. Gregory, R.L. Coleman (2003) "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data", J Biomed Biotechnol, 2003:308-314.

[17] J. Yu, X.W. Chen (2005), "Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data", Bioinformatics 2005, 21 Suppl 1:i487-i494.

[18] K. Ning, H.K. Ng, H.W. Leong (2006), "PepSOM: an algorithm for peptide identification by tandem mass spectrometry based on SOM", Genome Inform 2006, 17:194-205.

[19] J.S. Yu, et. al. (2005), "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data", Bioinformatics 2005, 21:2200-2209.

[20] B. Wu, T. Abbott, et. al. (2003), "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data", Bioinformatics 2003, 19:1636-1643.

[21] C. Burges.(1998) "A Tutorial on Support Vector Machines for Pattern Recognition". Data Mining and Knowledge Discovery 2:121 – 167.

[22] C. Cortes and V. Vapnik,(1995) "Support-Vector Networks", *Machine Learning, 20,*

[23] Van Looy *et al,(2007),"*A novel approach for prediction of tacrolimus blood concentration in liver transplantation patients in the intensive care unit through support vector regression.", *Critical Care,* **11**:R83

[24] Edda Leopold and Jörg Kindermann, (2006), "Content Classification of Multimedia Documents using Partitions of Low-Level Features", Journal of Virtual Reality and Broadcasting, no. 6, December 2006, urn:nbn:de:0009-6-7607, ISSN 1860-2037

[25] B. Schölkopf, A. Smola, and K. R. Muller, (1996) "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", Max-Planck-Institut für biologische Kybernetik, Arbeitsgruppe Bülthoff.