21st IEEE International Symposium on Computer-Based Medical Systems

An Integrated Clinico-Proteomics Information Management and Analysis Platform

M. Kalaitzakis[1], V. Kritsotakis[1], H. Kondylakis[1], G. Potamias[1], M. Tsiknakis[1], D. Kafetzopoulos[2]
[1]*Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Crete, Greece*
[2]*Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, Heraklion, Crete, Greece*
{mkalaitz, vkrits,kondylak,potamias,tsiknaki}@ics.forth.gr, kafetzo@imbb.forth.grl

Abstract

Detecting proteins in human blood holds the promise of a revolution in cancer diagnosis. Also, the ability to perform laboratory operations on small scales using miniaturized (lab-on-a-chip) devices has many benefits. Designing and fabricating such systems is extremely challenging, but physicists and engineers are beginning to construct such highly integrated and compact labs on chips with exciting functionality.

This paper focuses on the presentation of the information technology layer in such an integrated platform that has been developed in the LOCCANDIA project. LOCCANDIA ultimate objective is to develop an innovative nano-technology based (lab-on-a-chip) platform for the medical-proteomics field. The paper presents the main engineering aspects and the architecture of the Integrated Clinico-Proteomic Environment.

1. 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 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 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. 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.

Recent advances in MS-based proteomic technologies coupled with bioinformatics may revolutionize medical diagnosis and cancer screening. Mass spectrometry approaches [1] 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 [2]. The need for new and relatively simple devices to allow for the translation of these research results to clinical practice is urgent.

In this paper we focus on the presentation of the design of an integrated information management and analysis platform for a full-chain clinico-proteomic analysis.

2. Integrated Clinico-Proteomics Environment (ICPE) Architecture

Past experience has shown that multivariate analyses 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. The common format should include all proteome data, linked when possible with other characteristics obtained from fractionation together with the study protocol, sampling procedure and sample preparation, calibration, matching and quality control criteria, as well as clinical data. Such a database does not currently exist in the public domain.

Therefore, in the context of the LOCCANDIA project, such a database was designed and developed, by building on the outputs of existing standardization initiatives.

3. Software Engineering Methodology

During the project's initial phases we came across a number of obstacles, such as; the complexity of the domain, the phased implementation approach and the evolving requirements. As a result, a decision was taken to employ a user-driven iterative and incremental development (IID). Each iteration should be a self-contained mini-project composed of activities such as requirements analysis, design, programming, and test. User driven iterative development implied that the

1063-7125/08 $25.00 © 2008 IEEE
DOI 10.1109/CBMS.2008.48

choice of features for the next iteration came from the users. Likewise, test-driven development techniques were adopted in order to achieve the desired improvement and functionality of the system.

The aforementioned non-traditional software engineering techniques were adopted in the development of the LIMS platform because: a) the requirements were constantly evolving b) they were not clearly specified even during the development phase and c) unknown factors were constantly occurring since the complete workflow was unknown a priori.

4. Data analysis suite

The following modules are part of the LOCCANDIA Information Management System (LIMS), as they exchange information on demand and are to be used after the Liquid Chromatography module of the whole LOCCANDIA chain.

4.1. Data pre-processing and profile reconstruction

This module estimates the concentration of the unknown proteins. Its functionalities can be divided in two major operations: *Data pre-processing* - which is responsible for standardizing data and correcting the time delay in retention time among spectrograms [3] - and *profile reconstruction* - that is applied in order to quantify the proteins.

Pre-processing is achieved through a block matching algorithm, which does not use Dynamic Time Warping. The proposed method is applied on Liquid Chromatography-Mass Spectrometry "images".

Profile reconstruction is based on the inverse problem approach and relies on a functional model that associates the concentration profile of the unknown molecular with the measurements of the spectrograms. This approach is based on chemometric methods and thus has the advantage of increasing the sensitivity and robustness to noise and perturbations as it takes into account the whole signal.

4.2. Integrated Components

4.2.1 Protein / Peptide Identification (Phenyx)

This module is responsible for identifying the proteins and peptides that are not in the initial targeted panel but it is possible to be present in the measurement. Moreover, the whole process chain is complemented with the identification of the amino acid sequence of endogenous peptide candidates. This module is based on the Phenyx software, which is a software platform powered by Aldente algorithm (http://www.expasy.org/tools/aldente) in order to identify proteins from peptide mass fingerprinting data.. Phenyx supports all the common peak list file formats and provides a flexible and interactive Web based interface enabling data submission and validation of the requested results.

4.2.2. Visualization (MSight)

The visualization module provides a sophisticated way of visualizing mass spectrometer raw data, reconstructed profiles and any relevant information for data analysis. Implementation of this module is based on MSight (http://www.expasy.org/MSight) software platform (figure 1); an image analysis software for liquid chromatography mass spectrometry. MSight provides the 2-D representation, visual analysis and comparison of obtained datasets from protein or peptide separation combined to MS.

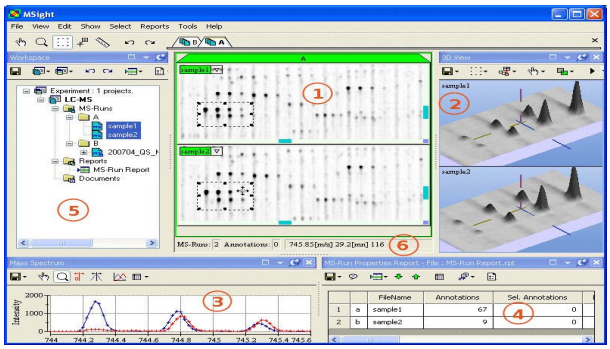

Figure 1. MSight view of related experiments

4.3. Knowledge Discovery and Data Mining

Mass spectrometry generates huge amounts of raw mass spectra, and analysis of these data may guide to the identification of known proteins, as well as the identification of novel ones. Moreover, coupling and analyzing such targets with real clinical data presents a promising direction towards the discovery of *biomarker* models. Biomarker models of *predictive* power being able to relate the various quantified protein concentration levels with different pathological conditions [2]. Proteomics approaches for the discovery of biomarkers are based on differential protein expression. Biomarkers need to be as *sensitive* as possible to avoid false negatives and also as specific as possible to avoid false positives. Utilization of advanced *knowledge discovery* and *data mining* technology is able to uncover potential 'hidden' patterns in the data and discover such models.

4.3.1. Utilizing Standard Data Mining Technology

In the course of the LOCCANDIA project we have assessed the reliability, efficacy and efficiency of various data mining approaches:

- R and R/**Bioconductor** (http://www.r-project.org): R is an open-source environment for statistical data analysis.. R/Bioconductor (http://www.bioconductor.org/) is an open source and open development software project, specially suited for the analysis and comprehension of genomic and proteomic data.

- MATLAB (http://www.mathworks.com/):

The MATLAB/**Bioinformatics Toolbox** provides access to genomic and proteomic data formats, analysis techniques, and specialized visualizations for proteomic analysis. The special MATLAB/BI-toolbox/**Mass Spectrometry Data Analysis** module provides set of functions for MS-data analysis: preprocessing, classification, and marker identification.

- Weka (http://www.cs.waikato.ac.nz/~ml/weka/):

Weka is a popular machine learning environment implemented in Java. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

- MineGene (www.ics.forth.gr/~kantale/MineGene.html) Developed by one of the LOCCANDIA partners, MineGene can serve a variety of supervised and unsupervised data mining operations. Initial experiments with MS-data have proved the utility of the system in proteomic data analysis as well.

5. LIMS

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 the aforementioned data analysis tools. Its ultimate objective is to intelligently correlate clinical and proteomic information towards LOCCANDIA's goal of early pancreatic cancer diagnosis.

5.1. Application workflow/System description

The authorized investigator, upon successful login, can access the LIMS application through its web-based graphical user interface (Figure 2). The application's interface consists of three main frames. The top frame displays information regarding the user's current actions while the left frame contains the application's navigational menu in a tree-like style format. The right, frame is dedicated to the action area of the application. This is where all the stored information is displayed. The user can navigate through the different entities of information, create new entries of data, insert additional information in already existing data and delete previous entered data. The interface is fully customizable and the user can display specific information based on defined search criteria.

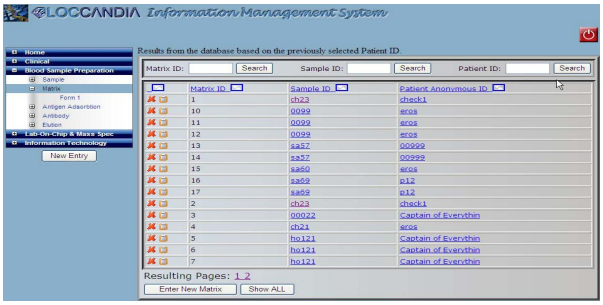

Figure 2. LIMS view information page

Information entry and manipulation of data is utilized through the use of forms. A link to each form is displayed on the application's navigation menu, organized according to the information entity they belong to.

6. 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.

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.

7. References

[1] Aebersold, R., and Mann, "M. Mass spectrometry-based proteomics". Nature, 422, 2003, pp. 198–207.

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

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