Prostate Cancer Disease Study by Integrating Peptides and Clinical Data

Patrizia Vizza [1] and **Luigi Pascuzzi** [2] and **Federica Aracri** [3] and **Elmiro Tavolaro** [4] and **Pasquale Lambardi** [5] and **Marco Gaspari** [6] and **Pietro Hiram Guzzi** [7] and **Giuseppe Tradigo** [8] and **Pierangelo Veltri** [9]

Abstract. Proteomic based analysis is used to identify biomarkers in blood samples and tissues. Data produced by devices such as Mass Spectrometry (MS), requires platforms aiming to identify and quantify proteins (or peptides). Clinical analysis can also be related with MS data.

In this work we focus on integrating clinical and biological data for prostate cancer in order to identify new biomarkers. We relate blood indicator (Prostate Specific Antigen, PSA) and urine samples analysis with MS based tissue analysis results. The focus is on relating tissue samples with neoplastic biomarkers [15]. The contribution proposes also a clinical data tool for tracking data and sample integrated with a tool box for information extraction.

1 Introduction

Studying chronic diseases data requires the collection and analysis of large amount of data (e.g., biological tissue sample and clinical data) [8, 23, 19]. The aim is to identify possible and useful biomarkers for the development of appropriate screening and prevention programs. A biomarker is an objectively measured characteristic describing a normal or abnormal biological state in an organism by analyzing biomolecules [11]. Cancer biomarkers are useful to measure the risk of developing cancer in a specific tissue, the risk of cancer progression or the potential response to therapy. Biomarkers can be classified into: (*i*) predictive biomarkers, which are able to predict responses to specific therapies, (*ii*) prognostic biomarkers, useful to estimate the risk of clinical outcomes, (*iii*) diagnostic biomarkers, used to identify whether a patient has a specific disease condition.

Databases and biobanks can be used in medical and biological research [17, 2, 3] to compare known available data and resources with measured ones. Biobanks allow the extraction, analysis and comparison of significant information, which can be used by domain experts as a support for the prevention or treatment of specific diseases. The set of biological samples (e.g. blood, biopsy tissues, body fluids) and patient's clinical information represent a fundamental tool to highlight molecular, genetic or environmental mechanisms and pathways in pathologies and to improve treatments in biomedical research [9], [5].

Even if prostate cancer (PCa) only affects men, it represents one of most diffused cancer in industrialized countries [13]. Prostate Specific Antigen (PSA) is the only biomarker widely used by physicians. Nevertheless it cannot be considered a reliable biomarker for its low specificity [7]. Thus, the identification of new biomarkers complementing or replacing PSA represents a main goal for prostate cancer research. MS-based biological sample analysis, as well as bioinformatics algorithms and statistics tools can support biomarker discovery research [10]. In literature, there are many approaches using bioinformatics and statistical algorithms in biomarker discovery which have been applied for accurate biological data analyses on patients [22, 1, 4]. A bioinformatic strategy for a quick identification of tissue-specific proteins, being also potential cancer serum biomarkers, has been proposed in [18]. In [21] the authors implement a clinical and biological database showing the utility of data integration to explore disease heterogeneity and to develop predictive biomarkers.

Authors in [26] identify lipid molecules useful for prostate cancer diagnosis by applying statistical methods as principal component analysis (PCA) and hierarchical clustering analysis (HCA) to analyze data.

In this paper we present the structure of an information system used to integrate information from clinical data and MS results regarding tissue and blood samples from patients affected by prostate disorders. The proposed system, which is a prototype for an ongoing research project, consists of a workflow manager able to track, store and analyze data obtained by monitoring patients who have been admitted in a clinical structure and provided biological sample to an MS laboratory.

The presented platform implements algorithms able to correlate clinical data (e.g. prostate gland dimensions) with peptides measures in a sample. Clinical data can also be correlated with demographic and environmental data stored in the platform's database.

The project's main goal was to identify a subset of interesting peptides through spectrographic analysis of blood serum, which represent natural biological markers significantly correlating with the presence or absence of prostate cancer. The implemented system, even if at an initial stage, is able to select interesting peptides which can be interesting candidate biomarkers for prostate cancer (PCa) and Benign Prostatic Hyperplasia (BPH).

[1] University of Catanzaro, vizzap@unicz.it
[2] VT Solution
[3] University of Catanzaro, federica.aracri@studenti.unicz.it
[4] VT Solution
[5] RelaTEch
[6] University of Catanzaro, gaspari@unicz.it
[7] University of Catanzaro, Italy, hguzzi@unicz.it
[8] eCampus University, giuseppe.tradigo@uniecampus.it
[9] University of Catanzaro, veltri@unicz.it

2 Clinical Data Tracking System

The proposed system integrates and analyzes clinical and molecular data in a single pipeline-based framework. Clinical analyses of prostate-related diseases are stored in a database and samples are precessed by MS analysis at Magna Graecia University laboratory with the goal of relating data and results for the identification of peptides as possible biomarkers in cancer prostate diagnosis.

A web based graphical user interface allows eased data entry and management. The web-based application architecture uses the Single Page pattern, implemented in Angular 6, where server modules have been implemented as a set of REST (Representational State Transfer) services, which store the status of the application on a MySql database instance.System architecture is shown in Figure 1.

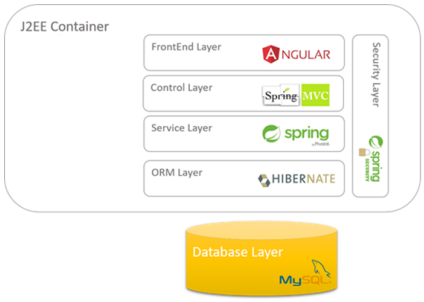

Figure 1. Platform architecture

2.1 Functionalities

The main system functionalities are: (i) data entry, (ii) tracking of patients in the clinical structures and (iii) tracking of blood and tissue samples. Information extracted from clinical database and from biological system have been anonymized in order to guarantee patients' privacy.Additional modules for data preprocessing, analysis and presentation have also been implemented: (i) statistic and analysis procedure definition module; (ii) dashboard for monitoring services and activities; (iii) data quality module; (iv) biological samples module, which retrieves from the database set of information for each sample (e.g. medical record number, recruitment date, age of patient, size of prostate gland); (iv) search module, able to retrieve biological samples or clinical information.

An example of data access and information extraction is reported in Figure 2.

The figure shows a list of biological samples. For each sample, a set of information are reported (e.g. medical record number, recruitment date, age of patient, size of prostate gland). *Sample* column reports the type of biological sample: it can be $blood$, $urine$ or both $blood\_urine$. *Biopsy Outcome* column expresses Gleason score of histologic exam.

3 Biomarker discovery process

Data analysis and mining algorithms implemented as modules of the presented platform, are able to take clinical and biological data stored in the platform's database and to identify specific peptides to be passed to a domain expert as potential biomarker for prostate cancer.

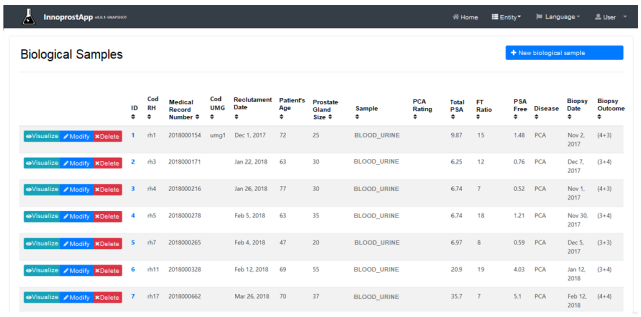

Figure 2. List of biological samples

Five different statistical algorithms have been included in the platform: (*i*) Pearson correlation coefficient [12], which measures linear correlation between two variables, X and Y, and it has a value between $+1$ and -1 for total positive and negative linear correlations respectively (values equal to 0 mean that there is no linear correlation between the two variables; (*ii*) Chi-square test, which is used to test the independence of two events [25]; given two variables, the test measures how observed count and expected count deviate from each other; when two variables are independent, the observed count is close to the expected count, resulting in a smaller Chi-square value (high Chi-square values indicate that the hypothesis of independence is incorrect); (*iii*) Recursive Feature Elimination (RFE) [14], used to fit a model and remove the weakest features thus eliminating existing colinearity by recursively eliminating features in an iterative process;(*iv*) LASSO (Least Absolute Shrinkage and Selection Operator) regression, which allows to automatically select variables [24, 16] in a high dimensional data space in order to perform regularization and variable selection; this could can improve both prediction accuracy and interpretation and works by minimizes the residual sum of squares providing that the sum of the absolute value of the coefficients being lower than a tuning parameter; (v) Finally, Random Forest (RF) algorithm has been implemented to classify PCA disease. RF is a combination of tree-structured predictors (decision trees) [20, 6], useful in molecular biology due to its flexibility and efficiency. RF can be used for a large number of predictor variables with limited sample sizes and genetic heterogeneity. Furthermore, the output tree is very useful for domain experts interpretation since it reports a decision tree with features thresholds generated by the algorithm to classify the objects in the dataset.

4 Results

The system has been implemented, tested and used to process and analyze data at the clinical structure partner of the project. Preliminary results on applying the algorithms implemented as modules of the system, which have been applied on almost 50 real cases, show interesting results in terms of: (*i*) possible interesting peptides that can be related with prostate cancer (i.e. novel biomarkers) and (*ii*) correlation among possible peptides and clinical data. The dataset contains a total of 54 patients, subdivided into 27 patients affected by PCA and 27 with BPH. Data resulting from biopsy and data extracted directly from the patient's medical record have been preprocessed as described above and stored on the database. Table 1 reports some of the main features including age, the size of the prostate gland (expressed as volume in *ml*) obtained by trans-rectal prostate ultrasound,

the value of Total PSA and Free PSA (both expressed in *mg/l*), and the ratio between Total and Free PSA (F/T Ratio). For each patient, a set of 32 peptides has been analyzed.

As a first experiment we implemented an ensemble-like approach according to which only the features satisfying at least 4 of the 5 algorithms have been considered. By using RF, we selected features (i.e. peptides) related to clinical information (e.g. age, dimension of prostate gland) in patients with PSA. Interesting peptides in terms of numerical and cluster results have been selected and are under consideration by clinicians.

5 Conclusion

Biomarker discovery represents an important task for the automatic discrimination of biological evidences in order to help domain experts in efficiently detecting prostate cancer at an early stage and in identifying aggressive tumors to improve patients care.

This paper describes a platform for the integration and analysis of clinical and molecular data. The platform provides modules able to identify possible biomarkers for prostate cancer identification.

ACKNOWLEDGEMENTS

This research has been supported by POR CALABRIA FESR-FSE 2014-2020 INNOPROST project. We are grateful to all colleagues working at the project and to Romolo Hospital as Reference of the project. Patrizia Vizza and Pierangelo Veltri are also supported by POR Telemetria 4.0.

REFERENCES

[1] D. Bonnel, R. Longuespee, J. Franck, M. Roudbaraki, P. Gosset, R. Day, M. Salzet, and I. Fournier, 'Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in maldi-msi: application to prostate cancer', *Analytical and bioanalytical chemistry*, **401**, 149–165, (2011).

[2] G. Canino, M. Cannataro, P. H. Guzzi, G. Tradigo, and P. Veltri, 'Relating clinical diagnosis and biological analytes via emrs clustering', *In 2014 IEEE International Conference on Healthcare Informatics*, 328–333, (2014).

[3] G. Canino, P. H. Guzzi, G. Tradigo, A. Zhang, and P. Veltri, 'On the analysis of diseases and their related geographical data', *EEE journal of biomedical and health informatics*, **21**, 228–237, (2015).

[4] M. Cannataro, G. Cuda, M. Gaspari, S. Greco, G. Tradigo, and P. Veltri, 'The eipeptidi tool: enhancing peptide discovery in icat-based lc ms/ms experiments', *BMC bioinformatics*, **8**, 255, (2007).

[5] Mario Cannataro, Giovanni Cuda, Marco Gaspari, Sergio Greco, Giuseppe Tradigo, and Pierangelo Veltri, 'The eipeptidi tool: enhancing peptide discovery in icat-based lc ms/ms experiments', *BMC bioinformatics*, **8**(1), 255, (2007).

[6] Young-Rae Cho, Marco Mina, Yanxin Lu, Nayoung Kwon, and Pietro H Guzzi, 'M-finder: Uncovering functionally associated proteins from interactome data integrated with go annotations', *Proteome science*, **11**(S1), S3, (2013).

[7] Y.A. Goo and D.R. Goodlett, 'Advances in proteomic prostate cancer biomarker discovery', *Journal of proteomic*, **73**, 1839–1850, (2010).

[8] N. Goossens, S. Nakagawa, X. Sun, and Y. Hoshida, 'Cancer biomarker discovery and validation', *Translational cancer research*, **4**, 256–269, (2015).

[9] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Giuseppe Tradigo, and Pierangelo Veltri, 'A time series approach for clustering mass spectrometry data', *Journal of Computational Science*, **3**(5), 344–355, (2012).

[10] A. Haoudi and H. Bensmail, 'Bioinformatics and data mining in proteomics', *Expert Review of Proteomics*, **3**, 333–343, (2006).

[11] P. Horvatovich and R. Bischoff, *Comprehensive Biomarker Discovery and Validation for Clinical Application*, Royal Society of Chemistry, 2013.

[12] H.C. Huang, S. Zheng, and Z. Zhao, 'Application of pearson correlation coefficient (pcc) and kolmogorov-smirnov distance (ksd) metrics to identify disease-specific biomarker genes', *BMC bioinformatics*, **11**, P23, (2010).

[13] A. Jemal, F. Bray, M.M. Center, J. Ferlay, and D. Ward, E. adn Forman, 'Global cancer statistics', *CA: A Cancer Journal for Clinicians*, **61**, 69–90, (2011).

[14] Y. Lv, Y. Wang, Y. Tan, W. Du, K. Liu, and H. Wang, 'Pancreatic cancer biomarker detection using recursive feature elimination based on support vector machine and large margin distribution machine', *4th International Conference on Systems and Informatics (ICSAI)*, 1450–1455, (2017).

[15] A Mazza, B Fruci, P Guzzi, B D'Orrico, R Malaguarnera, P Veltri, A Fava, and A Belfiore, 'In pcos patients the addition of low-dose spironolactone induces a more marked reduction of clinical and biochemical hyperandrogenism than metformin alone', *Nutrition, Metabolism and Cardiovascular Diseases*, **24**(2), 132–139, (2014).

[16] Giovanni Nassa, Roberta Tarallo, Pietro H Guzzi, Lorenzo Ferraro, Francesca Cirillo, Maria Ravo, Ernesto Nola, Marc Baumann, Tuula A Nyman, Mario Cannataro, et al., 'Comparative analysis of nuclear estrogen receptor alpha and beta interactomes in breast cancer cells', *Molecular BioSystems*, **7**(3), 667–676, (2011).

[17] S. Patil, B. Majumdar, K.H. Awan, G.S. Sarode, S.C. Sarode, A.M. Gadbail, and S. Gondivkar, 'Cancer oriented biobanks: A comprehensive review', *Oncology Reviews*, **12**, 357, (2018).

[18] I. Prassas, C.C. Chrystoja, S. Malawita, and E.P. Diamandis, 'Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery', *BMC Medicine*, **10**, 39, (2012).

[19] M. Prosperi, A. Pironti, F. Incardona, G. Tradigo, and M. Zazzi, 'Predicting human-immunodeficiency virus rebound after therapy initiation/switch using genetic, laboratory, and clinical data', *In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 611–516, (2016).

[20] M. Ram, A. Najafi, and M.T. Shakeri, 'Classification and biomarker genes selection for cancer gene expression data using random forest', *Iranian journal of pathology*, **12**, 339, (2017).

[21] M.D. Sorani, W.A. Ortmann, E.P. Bierwagen, and T.W. Behrens, 'Clinical and biological data integration for biomarker discovery.', *Drug discovery today*, **15**, 741–748, (2010).

[22] N.C. Tan, W.G. Fisher, K.P. Rosenblatt, and H.R. Garner, 'Application of multiple statistical tests to enhance mass spectrometry-based biomarker discovery', *BMC Bioinformatics*, **10**, 144, (2009).

[23] G. Tradigo, C. Veneziano, S. Greco, and P. Veltri, 'An architecture for integrating genetic and clinical data', *Procedia Computer Science*, **29**, 1959–1969, (2014).

[24] M.M. Vasquez, C. Hu, D.J. Roe, Z. Chen, M. Halonen, and S. Guerra, 'Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application', *BMC medical research methodology*, **16**, 154, (2016).

[25] L. Wang, Z. Jiang, M. Sui, J. Shen, C. Xu, and W. Fan, 'The potential biomarkers in predicting pathologic response of breast cancer to three different chemotherapy regimens: a case control study', *BMC cancer*, **9**, 226, (2009).

[26] X. Zhou, J. Mao, J. Ai, Y. Deng, M.R. Roth, C. Pound, J. Henegar, R. Welti, and SA. Bigler, 'Identification of plasma lipid biomarkers for prostate cancer by lipidomics and bioinformatics', *PloS one*, **7**, e48889, (2012).

Table 1. Dataset characteristics

	Age		Size of the prostate gland		Total PSA		Free PSA		F/T Ratio	
	PCA	BPH	PCA	BPH	PCA	BPH	PCA	BPH	PCA	BPH
mean	66	69	39.78	71.67	10.33	4.02	18.41	39.22	1.73	1.49
std	6.23	6.49	14.26	35.86	11.47	5.09	10.88	19.83	1.36	1,95
min	47	56	20.00	30.00	3.01	0.07	1.00	0.10	0.52	0.05
25%	63	66	30.00	50.00	6.11	0.91	14.00	23.50	0.98	0.20
50%	67	71	36.00	66.50	6.75	2.73	16.00	40.00	1.21	0.93
75%	72	73	48.25	83.25	8.35	4.49	21.00	54.50	1.68	2.10
max	77	81	75.00	173.00	58,40	21.86	62.00	79.00	5.65	9.43