33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Unifying the Integration, Analysis and Interpretation of multi-omic datasets: Exploration of the disease networks of Obstructive Nephropathy in children

Panagiotis Moulos, Ioannis Valavanis, *Member, IEEE*, Julie Klein, Ilias Maglogiannis, Senior *Member, IEEE*, Joost Schanstra, Aristotelis Chatziioannou, *Member, IEEE*

***Abstract*—The wealth of data amassed by the utilization of various high-throughput techniques, in various layers of molecular dissection, stresses the critical role of the unification of the computational methodologies applied in biological data handling, storage, analysis and visualization. In this article, a generic workflow is showcased in a multi-omic dataset that is used to study Obstructive Nephropathy (ON) in children, integrating microarray data from several biological layers (transcriptomic, post-transcriptomic, proteomic). The workflow exploits raw measurements and through several analytical stages (preprocessing, statistical and functional), which entail various parsing steps, reaches the visualization stage of the heterogeneous, broader, molecular interacting network derived. This network, where the interconnected entities are exploiting the knowledge stored in public repositories, represents a systems level interpretation of the pathological state probed.**

I. Introduction

RECENT advances in biological high-throughput techniques allow the simultaneous collection of data, derived from multiple sources of biological description. The microarray technology, widely used for probing gene expression, is now accommodating other levels of molecular dissection. Thus, nowadays microarrays are used to measure events at the post-transcriptional level of gene regulation (e.g. miRNA expression [1]), at the post-translational level by measuring active protein levels (antibody arrays [2]) or other like epigenomic (transcription factor binding [3]) and genetic events (SNP arrays [4]). In any case, the researcher is required to handle large amounts of data and complex output molecular lists, characterizing variegated experimental configurations [1].

Nowadays, the data management and analysis pipelines for data stemming from several layers of microarray analysis are gradually converging to standardized processing avenues, enabling the reproducible derivation of reliable biological outcomes, always under the premise of certain considerations [5]. However, rapid technological advances regarding Next Generation Sequencing, have a dramatic impact leading to the explosion of the throughput rates, concerning tasks such as de novo genome sequencing, re-sequencing, study of chromatin methylations and genome wide protein-DNA interactions [6]. The statistical challenges, regarding the analysis and meaningful interpretation of these entities, mark out the idiosyncratic nature and the limitations towards the goal of a systems level interpretation. This ultimately implies a failure of the interpretation regarding the understanding of the underlying molecular mechanisms concerning the interrogated system.

A big part of the bioinformatics routine represent tedious, yet extremely time-consuming, data-integration tasks from multiple sources, both experimental (e.g. different levels of transcription and translation events) as well as information stored in public biological databases and repositories. The former concerns quantitative information derived from different molecular biology experimental technologies, like gene expression (GEO), protein-protein interactions (IntAct), miRNA expression and detection (miRbase) for instance. The latter represents semantic information regarding different, levels of functional annotation in varying description depth, complying to the structure of ontological vocabularies (e.g. Gene Ontology) and model pathway maps (e.g. KEGG). Recently, there are approaches published, succeeding in integrating data through the use of relational database models, which bridge together a variety of public repositories (e.g. BRM [7]), by utilizing putative transcriptional networks [8], and graph visualization tools, in order to visualize the interaction landscape among different layers (genes, functions, diseases) [9].

However, the idiosyncrasies of multi-omic data integration stress the pluripotent nature of the interpretation tasks, leading to software development of numerous

Manuscript received April 14, 2011. This work was supported by Information Society Technology program of the European Commission "e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Sciences (e-LICO)" (IST-2007.4.4-231519).

P. Moulos is with the Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Athens, Greece (e-mail: pmoulos@eie.gr).

I. Valavanis is with the Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Athens, Greece (e-mail: ivalavan@eie.gr).

I. Maglogiannis is with the University of Central Greece, Lamia, Greece (e-mail: imaglo@ucg.gr).

J. Klein is with the Institut National de la Santé et de la Recherche Médicale (INSERM), U858, Toulouse, France and the Université Toulouse III Paul-Sabatier, Institut de Médecine Moléculaire de Rangueil, Equipe n°5, IFR150, Toulouse, France (e-mail: julie.klein@inserm.fr).

J. Schanstra is with the Institut National de la Santé et de la Recherche Médicale (INSERM), U858, Toulouse, France and the Université Toulouse III Paul-Sabatier, Institut de Médecine Moléculaire de Rangueil, Equipe n°5, IFR150, Toulouse, France (e-mail: joost-peter.schanstra@inserm.fr)

A. Chatziioannou is with the Institute of Biological Research and Biotechnology, National Hellenic Research Foundation, Athens, Greece, (corresponding author, phone: +30-210-7273751; fax: +30-210-7273758; e-mail: achatzi@eie.gr).

scattered, functionally disconnected amid each other, tools. Some exploit a complex database model or through the use of graph visualization libraries, visualize different layers of biological information (e.g. Cytoscape). Apart from the lack of a generalized integration approach, these solutions fail to incorporate the data analysis part, leaning to the expediency of other developed analytical tools [9], underrating the impact of the data-entry phase due to format inconsistencies. They also presume that the user is acquainted with high-throughput biological data analysis. Even those which integrate analysis modules together with biological database functionalities miss an analytical workflow that would enable labour-free comprehensive visualizations [10].

In this study, a generalized, framework for multi –omic analysis is presented, generic enough to handle various high-throughput data modalities (microarray data, next generation sequencing data). The resulting workflow encompasses versatile steps, like pre-processing, normalization, statistical, functional and pathway analysis for each type of data. It fuses the analytical tasks in a unified, comprehensive visualization of the multiple layers of information, through intense, intermediate parsing, enabling automated, flexible visual instantiation. In the use case, we integrate for demonstration purposes 3 –omics datasets that were used to study different severity grades of Obstructive Nephropathy (ON). ON is considered the most common children nephropathy, and the primary reason for kidney transplantations in children. It initiates as a renal pathological state caused by impaired flow of urine or tubular fluid [11], obstructing in the end-stage the proper urine flow. ON is of great importance to clinicians and common in infants due to congenital abnormalities of the urinary tract [12].

II. Materials And Methods

A. Data

The three –omics datasets analyzed setting the multi-omic dataset regarding the study of ON comprise i) human proteomics data ii) human miRNA data and iii) mice mRNA data. The human proteomic dataset included twenty children aged between two weeks and six months divided into four equal sized groups. Based on a set of clinical parameters, the samples were partitioned in three physiological subsets of five subjects each, namely: 1) Control including children without any renal damage, 2) No_Op comprising children with mild obstruction who do not need to undergo surgery to repair the ureteropelvic junction, and 3) Op: children with severe obstruction, who needed surgery to repair and reconstruct the junction. For each pediatric subject an antibody array allowed the quantification of the expression of 725 proteins in the collected urine. Human miRNA dataset encompassed infants divided, similarly with the proteomics dataset, into a Control subset (8 subjects), a No_Op subset (8 subjects) and an Op subset (10 subjects). The Agilent Human miRNA Microarray platform was used to measure expression values for a total number of 790 miRNAs. Three mice mRNA data groups were extracted after partial unilateral ureteral obstruction on neonatal mice in order to mimic the obstructive nephropathy syndrome as it occurs in children and examine the fingerprint of ON at the transcriptomic layer. These groups comprised non-operated Control mice (9), operated mice with Mild obstruction (5) and operated mice with Severe obstruction (5). Agilent's mice oligonucleotide microarrays were used to analyze the expression of 41000 mouse transcripts. Only the human homologues were considered.

B. Ontological and Pathway Analysis

For each dataset, over-represented GO terms and KEGG pathways were identified using the StRAnGER web application [13]. StRAnGER performs functional analysis of high-throughput genomic datasets, starting from a list of significant genes and using established statistical tests coupled with bootstrapping to derive a final population of statistically significant ontological terms.

C. Data Integration (KUPKB) and visualization

The integration of the multiple –omics entities was enabled through the use of the interactions stored in the Kidney and Urinary Pathway Knowledge Base (KUPKB). The KUPKB [14] uses Semantic Web technologies to integrate data and knowledge related to the Kidney and Urinary pathways, aiming to assist in biomarker discovery and molecular pathway modeling of diseases related to the urinary system. All the derived elements from the multiple –omics dataset analysis were visualized in EGAN [9].

III. Results

Statistical selection on the proteomics level revealed 43 Differentially Expressed (DE) proteins through the application of ANOVA among all subsets ($p<0.1$). From the DE proteins, over-representation of GO terms (GOTs) (11 in total) was observed targeting either general biological processes like protein binding, chromosome, DNA binding, nucleus or referring to more specific biological actions like regulation of cell cycle, spindle, which refers again to a specific phase of cell cycle, response to hyperoxia and proapoptotic caspase-mediated activities. At the gene expression level, the number of significant DE mouse transcripts obtained using a t-test ($p<0.01$, Benjamini-Hochberg FDR<0.25 and fold change cutoff $>|0.6|$ in $\log_2$ scale) were found 318 and 697 for the contrasts Control vs Mild and and Control vs Severe, respectively. Merging the DE mouse transcripts and then applying enrichment analysis revealed 99 over-represented GOTs and 22 KEGG pathways. Some of the GOTs reported here included several interesting functions related to inflammation and response to inflammation such as chemokine activity, oxidoreductase activity, oxidation reduction, growth factor activity, lipid metabolic process, cell migration and fatty acid metabolic process. On the miRNA level, a t-test ($p<0.01$) yielded 66 and 76 DE miRNAs for the contrasts Control vs No_Op and

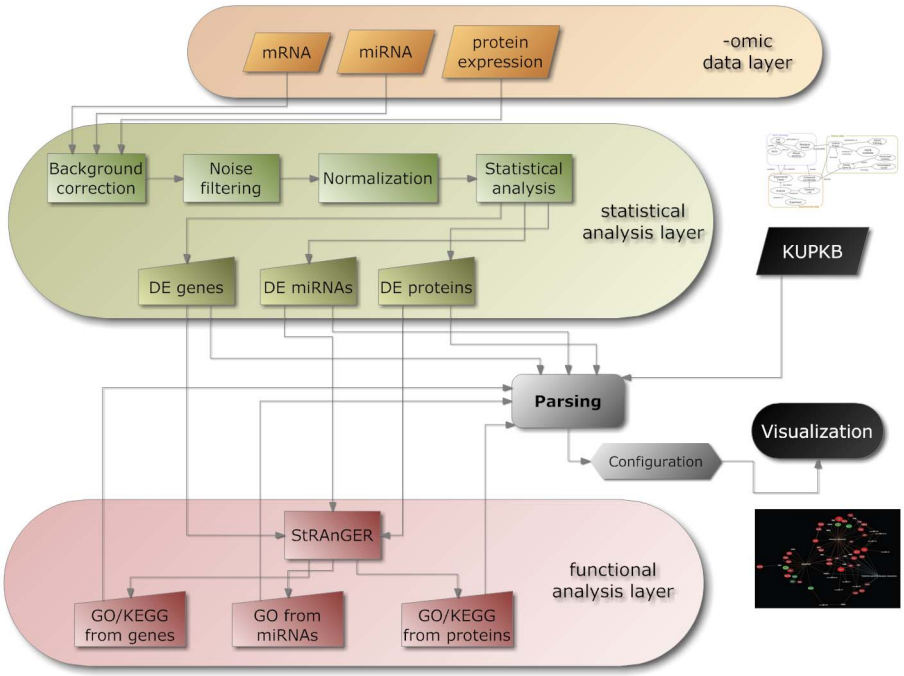

Fig. 1. The analysis workflow used to process the mutli –omics ON dataset. Starting from raw microarray image analysis data, the latter undergo several rounds of preprocessing, statistical and functional analysis, up to the point of visualization of a putative network derived by the analytical steps and data present in public repositories.

Control vs Op, respectively, while the enrichment analysis revealed 18 and 16 over-represented GOTs for the same contrasts using genes where the miRNA sequence is found within, and 111 and 101 GOTs using miRNA target genes based on interactions stored in the KUPKB.

As a next step, the visualization of the ON specific DE entities from the multi-omic analysis was endeavored in one generalized graph, which comprises putative as well as experimentally inferred relationships among molecular enities. For this scope, EGAN [9] was used, a Java application for gene-based graph visualization of high-throughput assay results, providing also a flexible interface to import several layers of customized interactions (node-node and associations). This interface creates visualizations with two possible ways i) direct node-node interactions and ii) node associations with super-nodes (e.g. pathways). Thus, the basic nodes, in this generalized graph, consist of genes derived from the gene expression analysis whereas the rest of the entities (miRNAs, proteins, GOTs, KEGG pathways) are depicted as super-nodes associated with multiple other entities. In this sense, the interconnections and possible interactions among the results of the different –omic layers (transcriptomic, proteomic, post-transcriptomic) are visualized in two levels: i) interactions among genes at the level of their expressed proteins using protein-protein interactions and ii) associations of genes (also proteins and miRNAs) with the respective biochemical pathways and GO functions they participate. miRNAs interact with the aforementioned entities in two ways: a) miRNAs that target genes derived from the transcriptomic data analysis and b) genes that encode for miRNAs derived from the post-transcriptomic (miRNA) data analysis.

For the visualization, the automated processing capabilities of Gene ARMADA [1] and StRAnGER [13] applications were exploited. The analysis results are then parsed by a parser exploiting several Perl scripts to accommodate several file types, providing additional options such as statistical filtering and keyword search. The parsing layer interfaces with the visualization software, in order to properly import the statistical and functional analysis results, taking into account additional options like the aforementioned keyword search. For example, the user is empowered to visualize only the GO terms “apoptosis” and “immune response” among a set of statistically significant GOTs derived by the functional analysis layer. Finally, a Perl script wraps all intermediate results and automatically creates a .jnlp file that is used to launch EGAN, with the customized visualization options. An overview of the workflow approach is presented in Fig. 1. All the interactions presented rely on the multi–omic data analysis, stored in the KUPKB for the following types of interactions: i) protein-protein interactions using UniProt IDs and mappings to HUGO gene symbols for *mus musculus* homologue genes in *homo sapiens* ii) predicted miRNA targeting genes for *homo sapiens* iii) genes coding for miRNAs in *homo sapiens* and iv) proteins associated to genes that express for those in *homo sapiens*. An instance of the network thus built is displayed in Fig. 2, depicting the GOTs “apoptosis” and “immune response”, the KEGG pathway “cytokine-cytokine receptor interaction”, the genes under these entities as well as the top twenty scoring miRNAs. From this figure, the researcher may for example observe that the majority of genes connected to “immune response” are intensely up-regulated, which is inline with the involvement of the immune system, depicted by earlier hypothesis or later findings, in the mechanisms beneath

nephropathy diseases [15-16].

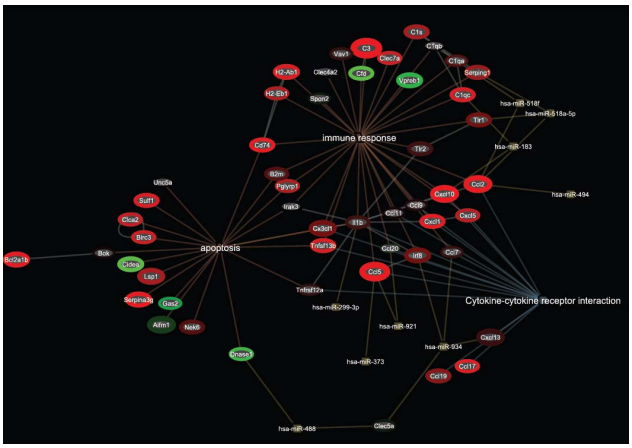

Fig. 2. A snapshot of the visualized network derived from the application of the workflow presented in Fig. 1.

IV. DISCUSSION

Despite the wealth of tools regarding various aspects of microarray analysis, unifying all the analytical steps easily and effectively remains a big challenge. In the past few years substantial effort has been deposited to the creation of analysis pipelines [17]. Regarding results visualization, even when the software infrastructure empowers it through the incorporation of several biological interaction databases [10], and the adoption of versatile visualization libraries (e.g. Cytoscape), the automation of the whole pipeline, from raw experimental data to network visualization, is missing. The innovation of the present work relies on the unification of all steps of multi-omic microarray datasets through several analytical layers within a single workflow in an automated fashion: the analysis starts from a common framework for the pre-processing and statistical analysis of raw data. Derived results are fed to a common functional analysis module and then to a dedicated parser, that can accommodate several file types including multi-omic entities (genes, miRNAs, proteins) and semantic entities of functional content (GOTs, KEGG pathways). The final result of the workflow is the reconstruction of a putative network based on current knowledge for interconnections of the aforementioned entities, with the use of software based on widely used graph visualization libraries [9]. The visualized network consisting of the most significant molecular entities and their interactions could help a biologist study in depth the mechanisms beneath a disease given that the appropriate experimental data is available.

We demonstrated our workflow based on data characterizing several experimental conditions regarding the severity of ON. The datasets used were derived from three different –omics layers, namely transcriptomic (gene expression), post-transcriptomic (miRNA expression) and proteomic (antibody microarrays to measure protein levels). The interactions among the DE entities were derived from a knowledge base built specifically to study the molecular mechanisms underlying the ON, the KUPKB. Future work includes among others the complete automation of the pipeline and the construction of a web interface which will provide users the opportunity to incorporate and process other datasets in a user-friendly manner.

REFERENCES

[1] A. Chatziioannou, P. Moulos, and F. N. Kolisis, "Gene ARMADA: an integrated multi-analysis platform for microarray data implemented in MATLAB," *BMC Bioinformatics,* vol. 10, pp. 354, 2009.

[2] R. Huang, W. Jiang, J. Yang *et al.*, "A biotin label-based antibody array for high-content profiling of protein expression," *Cancer Genomics Proteomics,* vol. 7, no. 3, pp. 129-41, May-Jun, 2010.

[3] S. Pillai, and S. P. Chellappan, "ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications," *Methods Mol Biol,* vol. 523, pp. 341-66, 2009.

[4] D. Nowak, S. Ogawa, M. Muschen *et al.*, "SNP array analysis of tyrosine kinase inhibitor-resistant chronic myeloid leukemia identifies heterogeneous secondary genomic alterations," *Blood,* vol. 115, no. 5, pp. 1049-53, Feb 4, 2010.

[5] S. Rosenfeld, "Do DNA microarrays tell the story of gene expression?," *Gene Regul Syst Bio,* vol. 4, pp. 61-73, 2010.

[6] W. J. Ansorge, "Next-generation DNA sequencing techniques," *N Biotechnol,* vol. 25, no. 4, pp. 195-203, Apr, 2009.

[7] A. R. Shah, M. Singhal, K. R. Klicker *et al.*, "Enabling high-throughput data management for systems biology: the Bioinformatics Resource Manager," *Bioinformatics,* vol. 23, no. 7, pp. 906-9, Apr 1, 2007.

[8] M. Bansal, V. Belcastro, A. Ambesi-Impiombato *et al.*, "How to infer gene networks from expression profiles," *Mol Syst Biol,* vol. 3, pp. 78, 2007.

[9] J. Paquette, and T. Tokuyasu, "EGAN: exploratory gene association networks," *Bioinformatics,* vol. 26, no. 2, pp. 285-6, Jan 15, 2010.

[10] S. Kozhenkov, Y. Dubinina, M. Sedova *et al.*, "BiologicalNetworks 2.0--an integrative view of genome biology data," *BMC Bioinformatics,* vol. 11, pp. 610, 2010.

[11] S. Klahr, "The geriatric patient with obstructive uropathy," *Geriatr Nephrol Urol,* vol. 9, no. 2, pp. 101-7, 1999.

[12] J. L. Bascands, and J. P. Schanstra, "Obstructive nephropathy: insights from genetically engineered animals," *Kidney Int,* vol. 68, no. 3, pp. 925-37, Sep, 2005.

[13] A. A. Chatziioannou, and P. Moulos, "Exploiting Statistical Methodologies and Controlled Vocabularies for Prioritized Functional Analysis of Genomic Experiments: the StRAnGER Web Application," *Front Neurosci,* vol. 5, pp. 8, 2011.

[14] S. Jupp, J. Klein, J. Schanstra *et al.*, "Developing a Kidney and Urinary Pathway Knowledge Base," in Bio-ontologies SIG, Boston, USA, 2010.

[15] D. Oliveira , "Membranous nephropathy: an IgG4-mediated disease" The Lancet, Vol. 351, 9103, pp. 670-671, 1998

[16] K. Hirayama *et al.*, "Predominance of type-2 immune response in idiopathic membranous nephropathy: Cytoplasmic cytokine analysis". Nephron, 91(2), pp. 255-61, 2002

[17] D. Montaner, J. Tarraga, J. Huerta-Cepas *et al.*, "Next station in microarray data analysis: GEPAS," *Nucleic Acids Res,* vol. 34, no. Web Server issue, pp. W486-91, Jul 1, 2006.