Integrating semantic transcriptomic data analysis and knowledge extraction from biological literature

Vid Podpečan*¶, Dragana Miljkovic*†, Marko Petek‡, Tjaša Stare‡, Kristina Gruden‡, Igor Mozetič*, Nada Lavrač*§
*Jožef Stefan Institute, Ljubljana, Slovenia
†Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
‡National Institute of Biology, Ljubljana, Slovenia
§University of Nova Gorica, Nova Gorica, Slovenia
¶Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia

***Abstract*—The paper presents an approach to the holistic analysis of transcriptomic data which integrates two state-of-the-art methodologies into a coherent framework. The aim of the proposed approach is to give insight into the discovered patterns, help explaining the observed phenomena, enable the creation of new research hypotheses and assist in design of new experiments. We have integrated a methodology for semantic analysis of transcriptomic data, a system for automated extraction of biological relations from the literature, and a number of supporting components. The approach is demonstrated and evaluated on a publicly available dataset from a clinical trial in acute lymphoblastic leukaemia and a document corpus of full-text articles from the PubMed Open Access Subset.**

***Keywords*-transcriptomic data; literature mining; data explanation;**

I. Introduction

Biological research and experiments are continuously producing new data and curated knowledge. However, similarly to other scientific disciplines our ability to analyse real-life experimental data in the context of already available curated knowledge has become limited. The enormous wealth of freely available biological knowledge requires multidisciplinary, integrative techniques that are able to overcome different formats, representations, encodings, and, most importantly, are able to reason and infer non-trivial knowledge which is hidden in the mass of data.

In this paper we propose an integrative approach, which merges two data analysis methodologies into a coherent framework. Our goal is to discover relevant knowledge in biological literature, which will explain and support the results of the analysis of transcriptomic experimental data. We have merged the SegMine methodology [1] for the semantic analysis of microarray data and the Bio3graph system [2], which implements automated extraction of biological relations from the literature. Both systems represent state-of-the-art development in the corresponding research field. One of our primary goals is also an easy-to-use implementation which allows repeatability of experiments and sharing of results. We aim to achieve this goal by developing the proposed approach, named **Seg3graph**, as a publicly available scientific workflow[1]. in ClowdFlows [3], a new generation data mining platform.

The proposed approach shares some of the design and data analysis features with few existing tools. One of the most well-known is Cytoscape [4], which integrates biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework. Cytoscape was used as a visualisation platform in the 3R (reading, reasoning, reporting) approach, implemented in the Hanalyser tool [5]. Our approach can be essentially classified as a 3R system although its reasoning components go beyond the capabilities of Hanalyser's reasoning processes as they implement specific state-of-the-art algorithms. In contrast with Hanalyser, reasoning is also performed on the experimental data.

The ONDEX system [6] aims to integrate multiple data sources, graph analysis and text mining to collect, present and analyse relevant data which can help to determine the biological relevance of detected significant changes in expression levels in the experimental microarray data. The text mining feature, however, is rather limited and ONDEX relies on external tools for the analysis of transcriptomic experimental data.

Tools such as Reactome [7], Biocyc [8], BioLayout [9] and MapMan [21] are examples of curated knowledge bases of metabolic reactions and pathways and computational tools to aid in the interpretation of microarrays and similar large-scale datasets. However, they are mostly limited to a few types of data where is left to the user to detect manually relevant connections. We refer the reader to the work of Leach et al. [5] for a discussion about few other systems which are related in terms of their architecture or the ability to integrate different background knowledge sources into the analysis of experimental data.

The rest of the paper is organised as follows. Section II presents the data sets used in this work and introduces the components of the presented approach. Section III discusses the results of the experiments, their evaluation and biological

[1]http://clowdflows.org/workflow/1129/

interpretation. Finally, Section IV summarises the work and points out the potential improvements and direction for further work.

II. Materials and methods

A. The data

1) Transcriptomic data: For the first part we have used a well-known microarray dataset from a clinical trial in acute lymphoblastic leukaemia (ALL) [10] which is a typical dataset for medical research. It contains 95 arrays for B-type cells and 33 arrays for T-type cells. We have used the results from [1] as a starting point. A union of enriched gene sets was computed which resulted in a set of exactly 100 genes which were then used in the second phase.

2) Document corpus: For the second part of our approach, where the goal is to extract biological knowledge from literature, we have obtained the relevant document corpus from PubMed Central Open Access subset (PMC OA) with the following query:

```
("t-lymphocytes"[MeSH Terms] OR
"t-lymphocytes"[All Fields] OR
"t cell"[All Fields] OR
"t-cell"[All Fields]) OR
("leukaemia"[All Fields] OR
"leukemia"[MeSH Terms] OR
"leukemia"[All Fields])
```

The query returned 77,168 publications while the full text was available for 62,142. These publications constitute our document corpus which was mined for specific biological knowledge related to the differentially expressed genes found during the first phase of our approach.

B. An overview of the Seg3graph approach

The approach presented in this paper fuses several processing components of two existing methodologies, SegMine [1] and Bio3graph [2] as well as extensive publicly available databases and search services into a novel data analysis workflow. Figure 1 depicts a schematic overview of the proposed approach.

Preprocessing of gene expression data. The goal of data preprocessing is to prepare the input for the discovery of differentially expressed gene sets. The data are recognised, parsed, validated, and the expression fold change is computed. A ranked list of genes is computed according to the observed importance in the raw data. Filtering options can also be applied to select a particular subset of genes.

SEGS algorithm. This component is one of the central elements of the workflow. The goal is to identify gene sets which are differentially expressed and which may reveal or indicate relevant underlying biological processes. Any known algorithm or method can be used for this task. The proposed approach uses the SEGS algorithm for this task as it offers the unique ability to describe the relevant

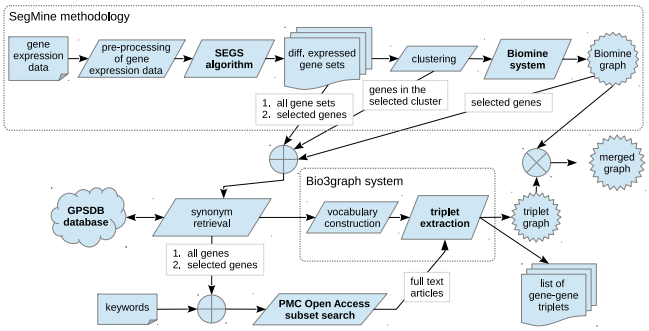

Figure 1. A schematic overview of the proposed Seg3graph approach. Note that the details of the triplet extraction in the Bio3graph system are not shown. Also, the results of triplet extraction typically require manual validation to remove false positives.

gene groups as logical rules formulated as conjunctions of ontology terms from GO, KEGG and Entrez. The rules semantically explain differentially expressed gene groups in terms of gene functions, components, processes, and pathways as annotated in biological ontologies. We refer the reader to the work by Trajkovski el al. [11] for a detailed description of the algorithm.

Clustering. As the SEGS algorithm may generate several similar rules, which can render the analysis difficult, clustering is used as a natural way of summarising the results. It also enables to focus on a selected cluster which can reduce the complexity of literature search and mining.

Biomine system. This component, which allows for discovery of new links, is based on the Biomine algorithm [12]. Biomine implements advanced probabilistic graph search algorithms on a graph structure merged from several public databases (Entrez Gene, UniProt, Gene Ontology, OMIM, NCBI HomoloGene, InterPro, Trembl, STRING, and KEGG). This component can be used in the proposed approach to uncover unexpected indirect relations and biological mechanisms. This knowledge can complement the biological relations discovered by Bio3graph and the results can be merged into a single graph structure and presented visually. We refer the reader to the work of Eronen and Toivonen [12] for a detailed discussion on Biomine.

Synonym retrieval. Components which retrieve and resolve gene synonyms are of crucial importance for the proposed approach. Despite existing guidelines many authors still describe genes and proteins using their own terms. Moreover, analyses of archived biological textual data are faced with even more confusing set of terms which may describe the same entity. We alleviate these problems by employing the Gene and Protein Synonyms DataBase (GPSDB) [13], which integrates 18 main up-to-date biological resources.

PMC Open Access subset search. Our data acquisition components are built around the *ESearch* and *EFetch* func-

tions provided by *E-utilies*[2]. The search component posts the query to *ESearch* and retrieves all identifiers of the relevant publications which are then processed by the XML parser which extracts the raw text.

Triplet extraction. The proposed approach relies on the Bio3graph tool to extract relevant, useful and compactly represented knowledge from textual sources. Bio3graph enables the extraction of knowledge (biological relations) in the form of triplets *(component1, reaction, component2)*. We refer the reader to the work of Miljkovic et al. [2] for a detailed description and evaluation of Bio3graph.

Vocabulary construction. We have retained the extensive Bio3graph's lists of reactions, their synonyms and passive forms while the list of components and their synonyms was constructed using our gene synonym retrieval component. In our experiments with the ALL dataset the mean value of synonyms per gene is 15.

III. Results and Discussion

For the first part of the Seg3graph analysis (identification of differentially expressed gene sets with SegMine) we have used existing, publicly available results [1] on the ALL dataset (see Section II-A and [10]). In total, 100 rules composed of ontology terms (GO, KEGG) and interaction terms (Entrez) describing differentially expressed gene sets were provided. Using the complete set of rules, all differentially expressed genes and all their synonyms were used to compose the Bio3graph component vocabulary (the Supporting Information S4 from [2] was reused as the reactions vocabulary).

In the second part of the analysis (processing of the literature), the Bio3graph system yielded 1.874 triplets (762 unique) from the literature dataset (see Section II-A for the literature dataset). They were then manually evaluated and 176 were found to be *correct* (true positive in at least one sentence). Among the correct triplets, biology experts recognised 113 as biologically sound. The achieved precision of Bio3graph in this use case is lower than expected [2] which can be attributed to the bigger and more general vocabulary.

Seg3graph allows for several different alternatives how the biological networks are obtained. Here, we have employed one of the more specific alternatives where the focus of the analysis is on one selected SEGS rule. We will perform a simple showcase analysis on the selected rule *GO:0030098 (lymphocyte differentiation)* $\wedge$ *Interact: KEGG:05340 (primary immunodeficiency)*. In total, this rule covers 16 differentially expressed genes in our data. The genes were used to construct a query to the Biomine system which discovers relevant links using several public databases and returns an annotated, weighted network.

In parallel, we have constructed a network using 113 biologically relevant triplets and performed a network merge

[2]http://www.ncbi.nlm.nih.gov/books/NBK25497/

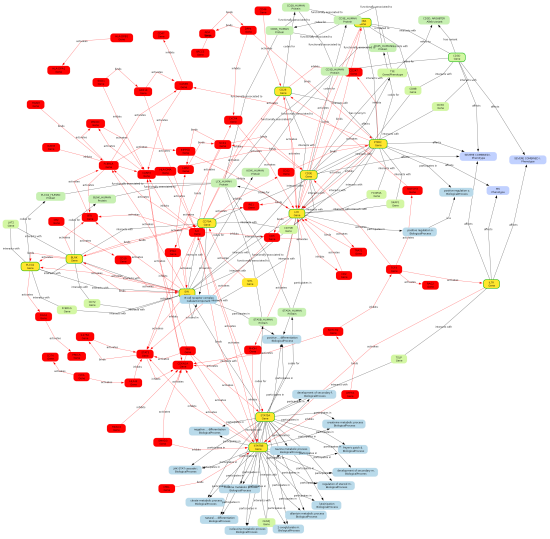

Figure 2. A biological network obtained by merging the Biomine network and the triplet network during the Seg3graph analysis of the rule *GO:0030098 (lymphocyte differentiation)* $\wedge$ *Interact: KEGG:05340 (primary immunodeficiency)*. Red colour denotes that the node or link was extracted from text. Yellow colour denotes that the node was discovered by Biomine but also extracted from text (other node colours are provided by Biomine and denote different entity types). Finally, green border denotes that a node was present in the Biomine query set (i.e., the node is a differentially expressed gene covered by the rule).

using different colours. The merged network is shown in Figure 2. It encodes several levels of biological information depending on the node and link colour. Figure 3 is a magnification of a small part of the network where the details of the colouring scheme can be observed.

The graph in Figure 2 contains 108 nodes and 213 relations in total, where 114 relations are extracted by Bio3graph and the rest (99 links) are found by the Biomine system. The relations between the yellow nodes usually confirm each other (the same knowledge is discovered by Biomine and Bio3graph), which enhances the confidence of domain experts in the obtained results. This is the case with the relations between BLNK and SYK gene (Figure 3). On the other hand, Bio3graph may extract a new direct relation between two genes (which was previously only

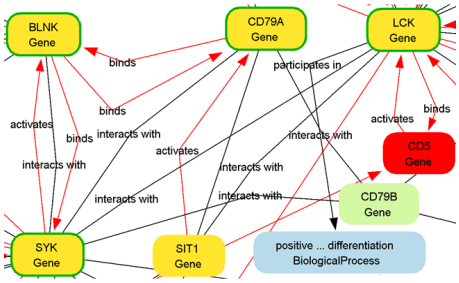

Figure 3. A magnification of a small part of the biological network from Figure 2.

indirect) that might be worth of further investigation. For example, it is known that the CD79A gene shown in Figure 3 initiates the activation of the signal cascade by binding of an antigen to the B-cell antigen receptor complex (BCR). The adaptor protein BLNK binds a signalling subunit of the BCR complex, and plays an important role in the BCR signal transduction [14]. Bio3graph has extracted from the literature a direct relation between BLNK and CD79A genes, while the Biomine system has found an indirect relation between these two genes through the SYK gene. More detailed biological investigation of the differentially expressed genes and their relations might bring interesting biological conclusions in the unified context of literature and experimental data.

IV. Conclusion

We have proposed and developed an approach to the holistic analysis of transcriptomic data by merging semantic analysis of experimental data and automated knowledge extraction from the literature. The proposed approach, named **Seg3graph**, was demonstrated on a dataset from a clinical trial in acute lymphoblastic leukaemia.

There are several directions for further work. Relation extraction from the literature can be improved to achieve greater accuracy, e.g., improvements proposed by Miljkovic et al. [2], recognition of mutants, etc. Our implementation does not yet fully support all the proposed options and additional components can be added to allow for even more complex scenarios. Support for different organisms is also planned. Finally, our evaluation was performed on a well-known (and relatively old) dataset. We plan to conduct an extensive evaluation on new experimental data using different supported analytical configurations.

Acknowledgment

This work was supported by the Slovenian Research Agency grants P4 0165, J4-2228, J4-4165, P2-0103, AD Futura scholarship and FP7 project MUSE (Machine Understanding for interactive StorytElling) under the Grant Agreement No. 296703.

References

[1] V. Podpečan, N. Lavrač, I. Mozetič, P. Kralj Novak, I. Trajkovski, L. Langohr, K. Kulovesi, H. Toivonen, M. Petek, H. Motaln, and K. Gruden, "Segmine workflows for semantic microarray data analysis in Orange4WS," *BMC Bioinformatics*, vol. 12, p. 416, 2011.

[2] D. Miljkovic, T. Stare, I. Mozetič, V. Podpečan, M. Petek, K. Witek, M. Dermastia, N. Lavrač, and K. Gruden, "Signalling network construction for modelling plant defence response," *PLoS ONE*, vol. 7, no. 12, p. e51822, 12 2012.

[3] J. Kranjc, V. Podpečan, and N. Lavrač, "Clowdflows: A cloud based scientific workflow platform." in *ECML/PKDD (2)*, ser. Lecture Notes in Computer Science, P. A. Flach, T. D. Bie, and N. Cristianini, Eds., vol. 7524. Springer, 2012, pp. 816–819.

[4] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research*, vol. 13, no. 11, pp. 2498–2504, nov 2003.

[5] S. M. Leach, H. Tipney, W. Feng, W. A. Baumgartner, Jr., P. Kasliwal, R. P. Schuyler, T. Williams, R. A. Spritz, and L. Hunter, "Biomedical discovery acceleration, with applications to craniofacial development," *PLoS Comput Biol*, vol. 5, no. 3, p. e1000215, 03 2009.

[6] J. Khler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Regg, C. Rawlings, P. Verrier, and S. Philippi, "Graph-based analysis and visualization of experimental results with ONDEX," *Bioinformatics*, vol. 22, no. 11, pp. 1383–1390, 2006.

[7] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, "Reactome knowledgebase of human biological pathways and processes," *Nucleic Acids Research*, vol. 37, no. Database-Issue, pp. 619–622, 2009.

[8] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. M. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp, "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 38, no. Database-Issue, pp. 473–479, 2010.

[9] A. Theocharidis, S. van Dongen, A. J. Enright, and T. C. Freeman, "Network visualization and analysis of gene expression data using BioLayout Express(3D)." *Nature protocols*, vol. 4, no. 10, pp. 1535–1550, 2009.

[10] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa, "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," *Blood*, vol. 103, no. 7, pp. 2771–2778, Apr. 2004.

[11] I. Trajkovski, N. Lavrač, and J. Tolar, "SEGS: Search for enriched gene sets in microarray data," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 588–601, 2008.

[12] L. Eronen and H. Toivonen, "Biomine: predicting links between biological entities using network models of heterogeneous databases," *BMC Bioinformatics*, vol. 13, p. 119, 2012.

[13] V. Pillet, M. Zehnder, A. K. Seewald, A.-L. Veuthey, and J. Petrak, "GPSDB: a new database for synonyms expansion of gene and protein names," *Bioinformatics*, vol. 21, no. 8, pp. 1743–1744, 2005.

[14] Y. Imamura, A. Oda, T. Katahira, K. Bundo, K. A. Pike, M. J. Ratcliffe, and D. Kitamura, "BLNK binds active H-Ras to promote B cell receptor-mediated capping and ERK activation," *J. Biol. Chem.*, vol. 284, no. 15, pp. 9804–9813, Apr 2009.