miXGENE tool for learning from heterogeneous gene expression data using prior knowledge

Matěj Holec, Valentin Gologuzov and Jiří Kléma
Department of Computer Science,
Czech Technical University,
Technická 2, Prague 6, 166 27, Czech Republic
Email: {holecmat,gologval,klema}@fel.cvut.cz

Abstract—**High-throughput genomic technologies have proved to be useful in the search for both genetic disease markers and more complex predictive and descriptive models. By the same token, it became obvious that accurate and interpretable models need to concern more than raw measurements taken at a single phase of gene expression. In order to reach a deeper understanding of the molecular nature of complexly orchestrated biological processes, all the available measurements and existing genomic knowledge need to be fused. In this paper, we introduce a tool for machine learning from heterogeneous gene expression data using prior knowledge. The tool is called miXGENE, it is elaborated upon in close connection with the biological departments that dispose of the above-mentioned data and have a strong interest in their integration within particular problem-oriented projects. The main idea is not merely to capture the transcriptional phase of gene expression quantified by the amount of messenger RNA (mRNA). The increasing availability of microRNA (miRNA) data asks for its concurrent analysis with the transcriptional data. Moreover, epigenetic data such as methylation measurements can help to explain unexpected transcriptional irregularities. miXGENE is an environment for building workflows that enable rapid prototyping of integrative molecular models.**

I. Introduction

During the last decade, the high-throughput genomic technologies like microarrays and next-generation sequencing have become an irreplaceable means for understanding genetically determined diseases [1]. Early research studies exploited gene expression (GE) data to discover sets of marker probesets without employment of existing (prior) knowledge [2]. This approach is not sufficient when an experiment results in too many relevant genes, which are hard to interpret, or too few genes, typically when an intricate interplay takes place among moderately expressed genes. Concerning complexity of the systems, volume of the data and the amount of noise within them, effective analysis of the data that takes into account existing knowledge has become a pressing need. It is now generally agreed that the true logic of diseases and other biological processes can only be explained by detailed interpretation [3].

There are a plethora of algorithms and tools designed to exploit prior knowledge (PK), typically based on the Gene Ontology (GO) terms [4] and cellular pathways (e.g., KEGG [5]), which provide information about the interplay of genes in various molecular functions, biological processes and cellular components, in the case of the GO, or provide information about molecular interactions in metabolic, signaling and disease-related pathways [6], [7], [8]). Still, several fundamental challenges of the enriched analysis using PK persist. This paper addresses the issue of moderate-only correlation between mRNA and protein levels which is caused by post-transcriptional and post-translational modifications [9]. These modifications not only play an important role in cell development and many diseases (e.g., [10], [11]) but also negatively affects the use of PK which is *not* directly based on transcriptional regulation interactions (e.g., protein-protein interaction networks) [12], [13]. This *lack of correlation* can be mitigated if we take into account other data sources; particularly, miRNA expression, since the small non-coding RNAs (miRNAs) play a role in translational and post-transcriptional regulation of GE and often result in gene silencing, and epigenetic data measuring level of DNA methylation and explaining unexpected transcriptional irregularities.

There are a variety of methods available for integrative analysis of mRNA, miRNA and methylation data and appropriate PK in the form of miRNA-mRNA interactions, cellular pathways and the GO terms; not all of them are capable to use all the data and knowledge. The most straightforward approach designed to exploit additional information in the form of the miRNA and methylation profiles is merging [14]; this method does not use any PK. [15] provides a black box integration procedure for several data sources with an immediate classification output. [16] utilizes matrix factorization, these methods employ PK such as miRNA-mRNA interactions and pathway definitions for regularization of the matrix decomposition. The methods described in [17], [10] infer miRNA and mRNA modules using miRNA-mRNA interactions based on decision rules and maximal bi-clusters, respectively. [18] uses GO terms and miRNA-mRNA relations to create ensemble classifiers based on tree-like modules. [19] proposes an integration method for mRNA and miRNA expressions directly motivated by the inhibition/degradation models of GE regulation.

This paper presents the web tool miXGENE freely available at `http://mixgene.felk.cvut.cz/`. It is designed mainly for joint enrichment analysis of mRNA, miRNA and DNA methylation data. miXGENE also contains an interface to the database repository of high throughput GE data (GEO) [20] and some other PK related databases (the GO [4], KEGG [5], MSigDB [21], miRWalk [22], miRBase [23]). miXGENE allows the user to create their own analytic pipeline using an interactive workflow editor and offers a spectrum of methods designed for data visualisation and analysis of mRNA, miRNA and DNA methylation profiles.

II. System description

miXGENE is representative of the mashup technology that fuses data from several publicly available sources (NCBI raw profiles and platform annotations, R Bioconductor libraries [24] and MSigDB [21]). The tool can be split into three parts: (i) graphical user interface (task definition, presentation of results), (ii) workflow management (task decomposition and its global planning in terms of the individual plugins) and (iii) computational plugins (implementation of the individual analytical methods such as data normalization, feature extraction, learning of classifiers, etc.). Web interface and storage management are implemented in the web application framework Django, workflow management is implemented in JavaScript and the computational plugins are mainly implemented in Python and R [25].

A. miXGENE as a workflow management system

miXGENE as a workflow management system are a growing area of research [26]. The main reasons for deployment of WMSs are: (i) an effort to make computational biology accessible to researchers who are not expert programmers, (ii) to enable tracking of experimental history and offer an easy-to-use tool for testing different settings, and (iii) the possibility to exchange the scientific workflows [26]. All these reasons are motivated by a goal to improve reproducibility, transparency and, therefore, mitigating experimental mistakes. There are many general frameworks or tools (both stand-alone and web-based) designed to represent bioinformatic or data-analytic workflows; e.g., BioBike [27], Taverna [28], Galaxy [29] and Anduril [30]. miXGENE can be seen as a specialized bioinformatics workflow management system. Despite the fact the mentioned WMSs (mainly the Galaxy) already implement some tools (several statistical test) and interfaces (GEO) we require, the WMSs are too general for our purposes. Therefore, in order to facilitate maintenance (e.g., keeping our system up-to-date and as specific for the joint analysis as possible), we implemented our own WMS.

With miXGENE, all experiments are built from components called blocks using interactive workspace. Each block represents one meaningful step in the experiment e.g., providing a source dataset, creating a ML model, visualisation. Nevertheless, each block usually contains—in contrast to the more general systems mentioned above—a few atomic activities such as downloading input data, preprocessing and diagnostic visualization in the *source dataset providing block*. The execution order is inferred from the data flow defined by binding the corresponding output and input ports of the consecutive blocks. miXGENE enables the block structured pattern [31] and it does not allow cycle dependency and conditional execution.

B. miXGENE building blocks and types

miXGENE defines two types of blocks: "basic“ *blocks* and *meta-blocks* where the latter serve as containers of other blocks or meta-blocks. The meta-blocks generate their own scope of possible input variables and, therefore, improve simplicity and clarity of workflows. This structure allows powerful and clear representation of ML workflows. Currently miXGENE enables blocks with the following functionality: (i) **data input** (access to GE data and knowledge from local user files or to selected public repositories), (ii) **data preprocessing** (tools for working with missing data and normalisation), (iii) **data manipulation** (simple data concatenation in case of compatible datasets; integrating different datasets, e.g., from different platforms measuring mRNA expression and joint analysis for data from mRNA, miRNA and methylation platforms; see Section III for details), (iv) **analysis** (various ML and statistical methods; see Section III for details), (v) **visualization** (results in human readable form, e.g., graphs, tables, textual descriptions of models, (vi) **performance evaluation (meta-block)** (evaluation schemes like k-fold cross validation or leave-one-out cross-validation), (vii) **multiple datasets evaluation (meta-block)** (for performing the same experiment on two or more datasets).

miXGENE operates with predefined complex data types rather than with a combination of atomic types like integer, string and array. Such an approach allows a required combination of data and meta-data for the desired level of workflow abstraction. E.g., the *Expression set* type contains GE data defined by a matrix dataset, phenotype description and platform annotation. Meta-data contains useful information about object content like data provider, used data type, properties of source tissue, etc. Data content is an object stored in fixed structure. Since data content may consume a great amount of memory, the complex data types allow serialization into the storage system.

List of implemented complex data types: (i) **expression set** (represents gene-expression data from a micro-array experiment including all necessary information), (ii) **gene set** (structure for representation of sets of genes, e.g. GO terms), (iii) **ML model** (learned model/classifier for the given data), (iv) **result table** (generic table in which each row represents features analysed during an experiment and each column represents different properties, metadata section contains a description of the column properties and working units), (v) **array container** (array of objects with the same structure, the cell structure description is stored in the metadata section).

C. Workflow construction

The main point of interaction between a user and the system is an experiment workspace with a block toolbox where the user defines an experimental workflow and executes it. The user constructs the new experiment from the empty workspace by adding appropriate blocks from the toolbox. To define data-flow, the user assigns input ports to outputs of the appropriate blocks. Then (if needed) the user sets mandatory or optional block parameters. When all the blocks in the experiment are configured correctly, the user can either execute each block by hand or run an automatic execution of the all blocks at the same time. The user will be notified about experiment's successful completion or will be pointed to occurred errors. The interactive nature of the experiment workspace allows the user to add more blocks anytime and continue the experiment with all the acquired results. Depiction of a machine-learning experiment based on a comparison of two alternative methods for analysis of mRNA and miRNA data is available via the miXGENE webpage. The shown workflow produces an estimate of accuracy of both methods and also final models based on the mentioned methods.

III. Methods

This section describes the methodological elements of our approach. The implemented WMS is primarily designed to support analysis using attribute-value machine-learning methods. These methods take input in the form of matrix where samples are in columns and features (e.g., probesets or genes) are in rows; each column contains a GE profile from one sample. In the case of supervised learning methods there is another vector with an assignment of each sample to a class of samples (e.g. healthy or cancerous tissue). The unsupervised learning methods do not include such a vector; instead, it makes its own classification using data properties. As input data, miXGENE currently supports a few human and mouse mRNA and miRNA platforms provided by Affymetrix and Illumina GoldenGate methylation assays. A complete list of the supported platforms is available on the miXGENE web.

A. Aggregating methods for knowledge enrichment

The aggregating methods (alternatively set-statistics methods) can incorporate PK in the form of gene-sets using a direct transformation which also produces matrix data representation. For example, there is the pathway p which is represented as a set of genes $g_1, g_2, \ldots, g_n$ and matrix with GE profiles where each row contains GEs for a gene g_j for the all samples. The aggregating methods transform the gene-expression matrix induced by the genes in the pathway p into a row vector which represents aggregated expressions for all the samples; such a vector is typically denoted by the name of the geneset p. The current miXGENE version supports the following methods: simple statistics as mean, median, PCA based transformation, and SetSig [7]. Thanks the flexible representation of workflows, the miXGENE does not impose any restriction on the gene-sets' definition; therefore, it is possible to use these aggregation functions anytime there are appropriate gene-sets which can define the transformation from the former to the new representation.

B. Data integration approaches

1) Integrated analysis of GE from different platforms and organisms: The integration is based on an assumption that it is generally possible to transform different data on the same common scale. For the integration of different MA platforms it can be mapping to the same genes and for different species it can be in evolutionary conserved elements like orthologous proteins. Generally, any common functionality describing gene sets like pathways or the GO terms can be used [32].

2) Joint analysis of mRNA miRNA and methylation profiles: Presently, miXGENE supports two joint-analysis approaches. The first one is the "naive" approach proposed in [14] which is implicitly accessible due to the flexibility of the workflow designer tool and power of the machine learning methods. It joins all of the types of datasets by columns; from the three datasets with mRNA, miRNA and methylation profiles which are represented by three matrices with features F_{miRNA}, F_{miRNA} and F_{methyl} the new "joint" dataset contains the set of features $F_{join} = F_{miRNA} \cup F_{miRNA} \cup F_{methyl}$. The second approach is based on a correction of mRNA expressions using miRNA expression profiles and known miRNA targets which describe the regulatory effect of miRNAs on mRNAs. miXGENE implements two versions of this approach; the substractive and the SVD-based method, which are suitable only for mRNA and miRNA data [19].

C. Other methods

miXGENE also implements other well established and state-of-the-art methods for analysis on single data with and without PK and on joint mRNA and miRNA datasets as referential standards. Different analytical approaches typically offer definite different solution due to the presence of alternative solutions (e.g., marker genes can point not on to disease causing genes but erroneously to genes related to a consequence of the disease) or unstable nature of the methods [33]. Moreover, the lack of gold standard data makes it impossible to compare alternative methods thoroughly; therefore, there is a need for the referential methods in order to problem being scrutinized to the depths necessary. For the PK-enriched analysis of mRNA expression, miXGENE integrates the global test [34]. As an alternative to the joint analysis methods we have implemented the algorithm based on generation context specific miRNA regulation modules based on GO terms [18].

IV. Case studies

Here we demonstrate miXGENE functionality in two biological case studies. A concise overview of results is available via the miXGENE webpage. The complete studies can be found in [13] and [19], respectively.

The first experiment focused on an evaluation of the hypothesis "gene set aggregation methods improve predictive accuracy if we use gene sets based on the structure of transcription regulation networks and on the operon structure of bacterial genomes" and was conducted solely on mRNA GE data. Recent studies reject this hypothesis for gene sets based on the GO terms and KEGG pathways [7], [12]. We evaluated this hypothesis on 71 small microarray GE datasets measured in the bacteria. The results on the bacterial data indicate that methods based on aggregation of gene sets are able to improve predictive accuracy when provided with suitable gene sets. When inappropriate gene sets are used, e.g., when one uses GO terms or KEGG pathways, then the accuracy may actually drop significantly.

In the second case study, we evaluated our novel feature extraction and data integration method for the accurate and interpretable classification of biological samples based on their mRNA and miRNA expression profiles. The main idea was to use the knowledge of miRNA targets and better approximate the actual protein amount synthesized in the sample. The raw mRNA and miRNA expression features become enriched or replaced by new aggregated features that model the mRNA-miRNA regulation instead. The underlying hypothesis is that "the sample profile presumably gets closer to the phenotype being predicted". The proposed subtractive aggregation method (SubAgg) directly implements a simple mRNA-miRNA interaction model in which mRNA expression is modified using the expression of its targeting miRNAs. This method works with the simplifying assumption of the equal weight of the individual miRNAs suitable for small sample sizes where learning of their proper weights may lead to overfiting. Its SVD-based modification (SVDAgg) enables different subtractive weights for different miRNAs learned by SVD. The two proposed knowledge-based subtractive methods were compared with their straightforward counterparts for obtaining the integrated mRNA and miRNA data through concatenating two respective datasets. We classified myelodysplastic syndrome patients under various experimental settings and compared the concatenation with SubAgg and SVDAgg. The results suggest that the knowledge-based approaches dominate the concatenation benchmark, and the features resulting from the mRNA-miRNA target relation can improve classification performance.

V. Conclusion

This paper presents a web tool for automated learning from heterogeneous genomic measurements that makes use of PK. The resulting models and markers match the actual measurements as well as the relationships among biological entities recorded in curated biological databases. The contribution of this tool is at least twofold. First, it provides the principal means for the user-friendly discovery of dedicated models in particular domains. Second, it is the platform for assembly, development, comparison and eventual dissemination of the methods for joint analysis of omics data. When compared with the traditional learning and statistical tools such as WEKA, RapidMiner, Orange or R/Bioconductor, it offers web interface with possibility to easily fetch NCBI data and implements specific learning methods, currently SubAgg and SVDAgg proposed in [19]. When compared with the bioinformatics WMSs such as Galaxy, it is focused on the specific task of learning from heterogeneous expression data. In particular, it facilitates the access both to the expression data and PK on their interaction, it provides specific learning methods and suggests sample workflows relevant to the given task.

Future work lies in further development and implementation of dedicated integration tools. We plan to continue with the development of our own methods as well as to employ the existing state-of-the-art algorithms. At the moment, there are no integration methods available for methylation and other epigenetic data available in miXGENE. We intend to improve miXGENE tool itself too, namely its graphical user interface and visualisation tools that serve for the presentation of results.

Acknowledgment

This research was supported by the Czech Republic Ministry of Health grant NT14539.

References

[1] K. Frese *et al.*, “Next-Generation Sequencing: From Understanding Biology to Personalized Medicine,” *Biology*, **1**: 378–398, 2013.

[2] T. R. Golub *et al.*, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” *Science*, **286**: 531–537, 1999.

[3] J. Dopazo, “Formulating and testing hypotheses in functional genomics,” *AI in Medicine*, **45**: 97–107, 2009.

[4] M. Ashburner *et al.*, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” *Nat Genet*, **25**: 25–29, 2000.

[5] M. Kanehisa *et al.*, “The KEGG resource for deciphering the genome,” *Nucleic Acids Res*, **32**: D277–280, 2004.

[6] P. Khatri *et al.*, “Ten years of pathway analysis: current approaches and outstanding challenges,” *PLoS Comput Biol*, **8**: e1002375, 2012.

[7] M. Mramor *et al.*, “On utility of gene set signatures in gene expression-based cancer class prediction,” *J Mach Learn Res*, **8**: 55–64, 2010.

[8] S. Hwang, “Comparison and evaluation of pathway-level aggregation methods of gene expression data,” *BMC Genomics*, **13**: S26, 2012.

[9] C. Vogel and E. M. Marcotte, “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses,” *Nat Rev Genet*, **13**: 227–232, 2012.

[10] X. Peng *et al.*, “Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers,” *BMC Genomics*, **10**: 373, 2009.

[11] M. Nugent, “MicroRNA function and dysregulation in bone tumors: the evidence to date,” *Cancer Manag Res*, **6**: 15–25, 2014.

[12] C. Staiger *et al.*, “Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis,” *Front Genet*, **4**, 2013.

[13] M. Holec *et al.*, “Gene-set features based on transcriptional regulatory network can improve simple gene-based classification,” unpublished.

[14] G. Lanza *et al.*, “mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer,” *Mol Cancer*, **6**: 54, 2007.

[15] D. Kim *et al.*, “Synergistic effect of different levels of genomic data for cancer clinical outcome prediction,” *J Biomed Inf*, **45**: 1191–8, 2012.

[16] S. Zhang *et al.*, “A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules,” *Bioinformatics*, **27**: i401–409, 2011.

[17] D. H. Tran *et al.*, “Finding microRNA regulatory modules in human genome using rule induction,” *BMC Bioinformatics*, **9**: S5, 2008.

[18] X. Zhou *et al.*, “Ensemble classifier based on context specific miRNA regulation modules: a new method for cancer outcome prediction,” *BMC Bioinformatics*, **14**: S6, 2013.

[19] J. Klema *et al.*, “Knowledge-Based Subtractive Integration of mRNA and miRNA Expression Profiles to Differentiate Myelodysplastic Syndrome,” in *Proc. Int. Conf. on Bioinform. Models, Methods and Algorithms*, 31–39, 2014.

[20] T. Barrett *et al.*, “NCBI GEO: archive for functional genomics data sets–update,” *Nucleic Acids Res*, **41**: D991–995, 2013.

[21] A. Subramanian *et al.*, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,” *PNAS*, **102**: 545–550, 2005.

[22] H. Dweep *et al.*, “miRWalk–database: prediction of possible miRNA binding sites by ”walking” the genes of three genomes,” *J Biomed Inform*, **44**: 839–847, 2011.

[23] A. Kozomara *et al.*, “miRBase: integrating microRNA annotation and deep-sequencing data,” *Nucleic Acids Res*, **39**: D152–157, 2011.

[24] R. Gentleman *et al.*, “Bioconductor: Open software development for computational biology and bioinformatics,” *Gen Biol*, **5**: R80, 2004.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[26] A. Barker and J. V. Hemert, “Scientific workflow: a survey and research directions,” in *Parallel Processing and Applied Mathematics*, ser. LNCS, R. Wyrzykowski *et al.*, Eds., 2008, **4967**: 746–753.

[27] J. Elhai *et al.*, “BioBIKE: a Web-based, programmable, integrated biological knowledge base,” *Nucleic Acids Res*, **37**: W28–32, 2009.

[28] K. Wolstencroft *et al.*, “The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud,” *Nucleic Acids Res*, **41**: W557–561, 2013.

[29] J. Goecks *et al.*, “Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences,” *Genome Biol*, **11**: R86, 2010.

[30] K. Ovaska *et al.*, “Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme,” *Genome medicine*, **2**: 65, 2010.

[31] M. La Rosa *et al.*, “Managing Process Model Complexity Via Abstract Syntax Modifications,” *IEEE T Ind Inform*, **7**: 614–629, 2011.

[32] M. Holec *et al.*, “Integrating Multiple-Platform Expression Data through Gene Set Features,” in *ISBRA*, 2009.

[33] S. Michiels *et al.*, “Prediction of cancer outcome with microarrays: a multiple random validation strategy,” *Lancet*, **365**: 488–492, 2005.

[34] J. J. Goeman *et al.*, “A global test for groups of genes: testing association with a clinical outcome,” *Bioinformatics*, **20**:93–99, 2004.