Cross-genome knowledge-based expression data fusion

Matěj Holec, Jiří Kléma, Filip Železný, Jiří Bělohradský
Czech Technical University,
Technická 2, Prague 6, 166 27
{holecm1,klema,zelezny}@fel.cvut.cz

Jakub Tolar
University of Minnesota, Minneapolis
tolar003@umn.edu

Abstract

This paper presents the web tool XGENE.ORG which facilitates the integration of gene expression measurements with background genomic information, in particular the gene ontology and KEGG pathways. The novelty of the proposed data fusion is in the introduction of working units at different levels of generality acting as sample features, replacing the commonly used gene units, consequently allowing for cross-genome (multi-platform) expression data analysis. The integration of different microarray platforms contributes to the robustness of knowledge extracted when single-platform samples are rare and facilitates inference of biological knowledge not constrained to single organisms.

1. Introduction

In the current post-genomic era, various aspects of gene functions are being uncovered by a large number of experiments producing huge amounts of heterogeneous data at an accelerating pace. Putting all this data together, while taking into account existing knowledge has become a pressing need for developing tools able to explore and simulate biological entities at a system level. A popular example is the microarray (MA) technology enabling to simultaneously estimate the activity of tens of thousands genes (virtually the entire genome) in a sample tissue. Early research studies exploited gene expression data to discover sets of marker probesets, e.g. those with elevated expression in a cancerous tissue. Despite several successes in predictive diagnosis using such obtained knowledge, it is now generally agreed that the true logic of diseases and other biological processes can only be explained by detailed interpretation of the measurements, clarifying how and why certain genes follow certain expression patterns in certain situations. This in turn requires to integrate the large volumes of raw measurements with another huge body of available additional information (or background knowledge, BK), such as known gene functions, mutual interactions or roles in regulatory and signalling pathways.

The most popular and frequent utilization of background knowledge is based on enrichment analysis. The state-of-the-art tools such as DAVID [7] search for enriched apriori-defined *gene groups*, rather than interpret individual differentially expressed probesets (or genes[1]). The principal foundation of enrichment analysis is that if a biological process is abnormal, the co-functioning genes should have a higher (enriched) potential to be selected as relevant. Such a rationale can move the analysis from an individual gene-oriented to a relevant gene group-based one. The overview of 68 available enrichment tools is available in [8]. The biological utility of pathways was demonstrated by the study [11] where a significantly downregulated pathway-based gene set in a class of type 2 diabetes was discovered despite no single significant gene being detected. [10] provides a method that uses gene ontology terms and their grouping to improve the interpretation of gene set enrichment for microarray data.

This paper presents the web tool XGENE.ORG available at `http://xgene.org`. Similarly to enrichment tools, XGENE.ORG tool facilitates integration of large volumes of raw gene expression measurements with another huge body of available genomic information. Contrary to existing enrichment tools, it offers additional functionality resulting from a data-fusion strategy based apriori defined gene sets. In particular, the main resulting feature of the present tool is that it enables to analyze gene expression data collected from heterogeneous platforms in an integrated manner. The heterogeneous platforms may pertain to different organism species. The significance of this contribution is at least twofold. First, microarray experiments are costly, often resulting in numbers of samples insufficient for reliable modeling. The possibility of systematically integrating the experimenter's data with numerous public expression samples coming from heterogeneous platforms, would obviously help the experimenter. Second, such in-

[1] In this paper we consider probesets and genes as closely related but still distinct units as several probesets may interrogate the same gene.

tegrated analysis provides the principal means to discover biological markers shared by different-genome species.

XGENE.ORG explicitly implements various *working units* and determines their *level of activity*. The activity of a superior (more abstract) working unit is calculated from the known (measured) activity of a set of inferior (less general) working units. For example, it selects all the probesets that are annotated by the same gene identifier and computes gene activity. Likewise, all the genes whose products act in a single pathway are used to compute pathway activity. A similar approach that applies a method based on singular value decomposition to calculate pathway activity was proposed in [19]. However, XGENE.ORG takes a step forward. First, it works with a various types of working units on different levels of generality. Second, it uses them to perform cross-genome and cross-organism analysis as there are working units that generalize beyond individual platforms and species. Third, in addition to standard statistical analyses, it applies machine learning (ML) techniques to develop interpretable models that distinguish among user-defined classes.

Let us exemplify some of the currently available types of working units. The first type that enables cross-platform analysis aggregates measurements that share a common gene ontology (GO) [3] term. The second type aggregates measurement units acting in the same biological pathways formalized by the KEGG [9] database. The third type represents a further novel contribution of our work and is based on the notion of a *fully coupled flux*, which is a pattern prescribing pathway partitions hypothesized by [12] to involve strongly co-expressed genes.

To sum up, analyses and models based solely on *measurement units* defined by the individual probesets whose expression is immediately measured by microarrays suffer from the inherent microarray noise and often fail to identify subtle patterns, give a large room to overfitting and prove hard to interpret and apply. Genomic background knowledge makes it possible to introduce and analyze alternative working units that avoid the bottlenecks mentioned above and provide improved interpretation power and statistical significance of analysis results. At the same time, different platforms and/or species deal with different sets of measurement units that cannot be directly matched. Consequently, multi-platform analyses cannot be performed without working units whose meaning is general enough to be defined in each platform and whose activity can unambiguously be evaluated in each sample independently of its platform type. Working units then serve as markers (or features) to distinguish between user-supplied sample classes.

The paper is organized as follows. In Section 2 we synthetize the system's functionality and describe its architecture. Section 3 describes the methodological elements of our approach, consisting of normalization, extraction of working units at various levels of generality, testing their significance, and predictive classification. Section 4 briefly exemplifies the use of the system through two case studies. Section 5 lays out prospects for future work and concludes the paper.

2. System Description

The main goal of the presented XGENE.ORG tool is to analyse a wide range of publicly accessible heterogeneous gene expression samples. The tool provides an interface to search available measurements whose annotation is relevant to the studied biological topic. Typically, a set of relevant measurements straddles various microarray platforms and organisms. There are two principal reasons to allow for their integration. The technical reason concerns the sufficiency of sample sets for reliable modeling. The more platforms accessed, the larger number of samples is at hand. The scientific reason pertains to the relevance of the outcomes. Combining multi-platform input data contributes to the generality of any knowledge discovered.

The tool operates in three basic phases:

1. define sample classes of interest; search and collect existing measurements representing these classes,
2. compute the activation levels of various working units with respect to the collected samples,
3. apply statistical, machine learning and visualization methods to obtain models distinguishing between the defined classes, with the pre-computed activity levels of working units acting as sample features.

XGENE.ORG implements this workflow, facilitating all three phases above. The architecture of the tool is depicted in Figure 1. XGENE.ORG integrates data from several publicly accessible databases.

Regarding the first phase above, our tool provides an interface to the Gene Expression Omnibus (GEO) [1]. XGENE.ORG enables a keyword-based search and filtering of individual gene expression measurements as illustrated in Fig 2. GEO is currently the largest public repository archiving and freely distributing high-throughput gene expression measurement data submitted by the scientific community. GEO currently stores approximately a billion individual gene expression measurements, derived from over 100 organisms, addressing a wide range of biological issues. GEO is accessible at www.ncbi.nlm.nih.gov/geo. The interaction with GEO is supervised by the user. The measurements are normalized and saved in the internal PROLOG format that simplifies subsequent integration of the expression data with data capturing biological process structure (pathways) and relational information (the gene ontology).

Secondly, XGENE.ORG accesses the databases that provide background knowledge required to define and interpret the predefined set of working unit types (they are discussed in detail thereunder). The individual microarray platforms are annotated by the Bioconductor packages [4]. Bioconductor packages also provide annotations by the gene ontology terms. The background knowledge on pathways and fluxes is taken directly from KEGG [9] database. The background knowledge management is fully automated and carried out without user interventions. The tool downloads all the packages and datasets needed to analyse the measurements currently selected by the user and stores them in the internal PROLOG representation.

The critical step is to fuse the collected measurements and background knowledge into unified cross-platform data subsequently accessed by the statistical and machine learning tools. Within this fusion, working units are computed across samples taken from various platforms and organisms. The resulting unified representation consists of a single matrix in which rows correspond to samples, columns correspond to working units and the respective matrix cells express the activity of a given unit within a given sample as a real value. Each working unit subsequently serves as a statistical variable for tasks such as fold change analysis, or a *sample feature* for machine learning algorithms.

Currently, three kinds of analysis results are supported:

- a classifier that estimates the sample class given an expression sample and its platform label
- a list of working units significantly differentially expressed in classes
- a scatterplot that shows class distribution in a (transformed 2D) space of working units.

The results are provided to the user in the form of hypertext, including links pointing to detailed descriptions working units employed in the displayed result.

The interaction with the user who starts a new experiment consists of the following steps:

1. The user logs to his/her personal account. This account stores the user's previous experiments and their results.
2. The user creates a new experiment. The experiment can be entirely new (the interaction proceeds by the following step) or it can be derived from a previous experiment (the experiment then inherits the classes and datasets defined earlier and thus skips the two following steps).
3. The user creates and entitles two or more of sample classes. These classes contain no measurement samples at this stage.
4. The user fills each of the defined classes with a set of relevant GEO expression samples. The samples are preselected via keyword-based search and then finely filtered by the user on the basis of experimental annotations (see Figure 2),

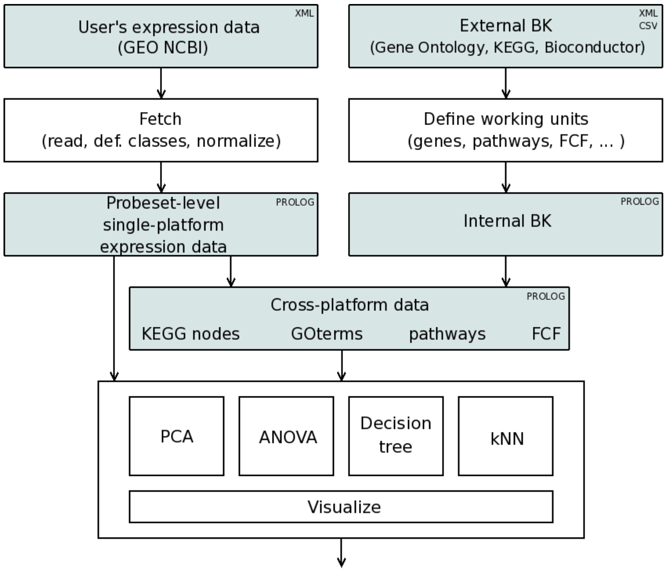
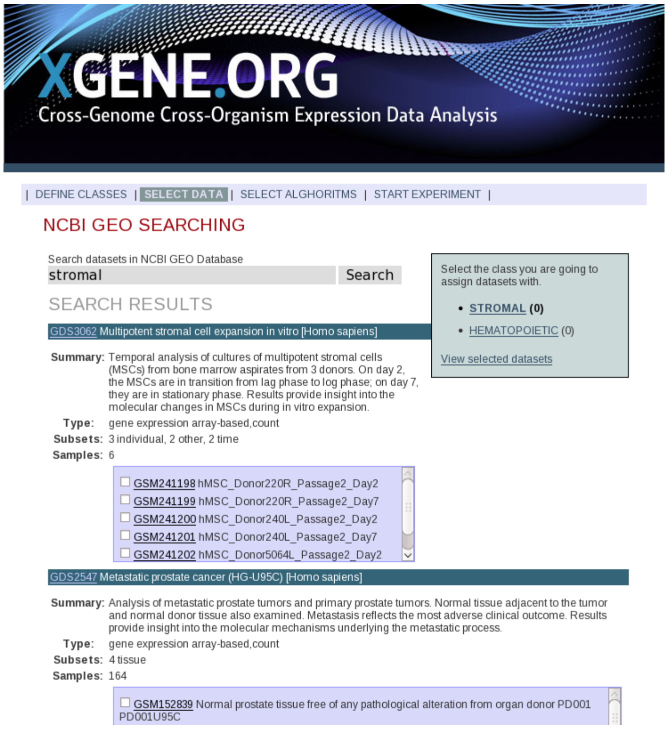


Figure 1. XGENE.ORG architecture

5. The user selects (possibly repeatedly) proper working units, platform types and algorithms and starts the experiment.
6. The system collects the necessary background knowledge, computes the working units defined above and applies the selected algorithms.
7. The computation begins and the user can log out. (S)he is informed by email as soon as the results are ready to be shown.
8. The user views the results. A result-filter helps user's orientation if a large number of result types has been requested in step 5.

3. Methods

This section describes the methodological elements of our approach. It gives an overview of working units and shows the way in which their activity is estinated and evaluated. It specifies the statistical methods serving to identify differentially expressed working units. It also gives a summary of currently implemented machine learning methods. Their application is at least twofold. The first one is practical. They provide means to distinguish among sample classes when the sample annotation is unknown. The second one is exploratory. As one of the keynotes of XGENE.ORG is to prove applicability of cross-platform working units, the *classification accuracy* of machine learning models is instrumental for relevance assessment of a given set of working units.


XGENE.ORG
Cross-Genome Cross-Organism Expression Data Analysis
| DEFINE CLASSES | SELECT DATA | SELECT ALGHORITMS | START EXPERIMENT |

NCBI GEO SEARCHING

Search datasets in NCBI GEO Database

stromal Search

SEARCH RESULTS

GDS3062 Multipotent stromal cell expansion in vitro [Homo sapiens]

Summary: Temporal analysis of cultures of multipotent stromal cells (MSCs) from bone marrow aspirates from 3 donors. On day 2, the MSCs are in transition from lag phase to log phase; on day 7, they are in stationary phase. Results provide insight into the molecular changes in MSCs during in vitro expansion.
Type: gene expression array-based,count
Subsets: 3 individual, 2 other, 2 time
Samples: 6

□ GSM241198 hMSC_Donor220R_Passage2_Day2
□ GSM241199 hMSC_Donor220R_Passage2_Day7
□ GSM241200 hMSC_Donor240L_Passage2_Day2
□ GSM241201 hMSC_Donor240L_Passage2_Day7
□ GSM241202 hMSC_Donor5064L_Passage2_Day2

Select the class you are going to assign datasets with.

- **STROMAL (0)**
- HEMATOPOIETIC (0)

View selected datasets

GDS2547 Metastatic prostate cancer (HG-U95C) [Homo sapiens]

Summary: Analysis of metastatic prostate tumors and primary prostate tumors. Normal tissue adjacent to the tumor and normal donor tissue also examined. Metastasis reflects the most adverse clinical outcome. Results provide insight into the molecular mechanisms underlying the metastatic process.
Type: gene expression array-based,count
Subsets: 4 tissue
Samples: 164

□ GSM152839 Normal prostate tissue free of any pathological alteration from organ donor PD001 PD001U95C

Figure 2. XGENE.ORG: collecting relevant samples from NCBI GEO. Clicking on a sample identifier ('GSMxxxxx') opens a detailed description of that sample.

3.1. Working units – types and activity

Currently, we consider two principal knowledge sources in order to define working units—the gene ontology database [3] and the KEGG database [9]. The Bioconductor annotation packages [4] serve to translate among the identifiers used by the microarray manufacturers (currently, only Affymetrix is supported), and the two mentioned background knowledge databases. The widely spread EntrezIds (gene identifiers) introduced by NCBI play the role of intermediate translation identifiers. The current hierarchy of working units as implemented in XGENE.ORG is shown in Figure 3. The ultimate working units correspond to the measurement units, i.e., the probesets. Their activity in the individual samples is directly reported in the GEO input files. A single GEO file corresponds to a single microarray sample, a whole sample is represented by a probeset activity vector. The set of measured probesets is platform dependent, i.e., the vectors taken form different platforms cannot be directly matched. The more general units are gradually inferred from their subordinate units. For example, the list of probesets that are annotated by the same gene identifier makes up the *gene* working unit. The list of genes linked to a pathway node makes up the *pathway node* working unit. To compute the activity of a working unit, the probesets that transitively link to that working unit are considered. For example, the activity of a pathway is computed by aggregating the activity of all probesets corresponding to genes which in turn correspond to nodes contained in the given pathway. Obviously, this mapping is platform dependent; pathways have different probeset interpretations in different platforms. At the same time, this mapping is organism dependent and thus we have to deal with organism orthologs of pathways.

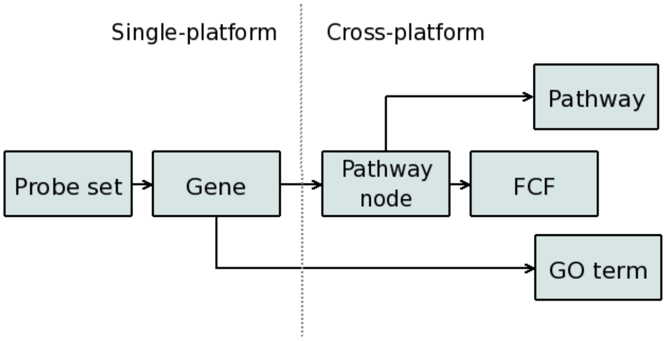


Figure 3. The hierarchy of working units. An arrow from X to Y denotes that unit Y refers to a set of X units. This relation is transitive and thus all units can ultimately be represented as families of probesets.

Significance testing at the level of pathways and/or GO terms is a standard method widely implemented in enrichment tools. However, these working units may prove overly general to capture subtle biological dependencies. Many notable biological conditions are characterized by the activation of only certain parts of pathways; for example, see references [16, 20, 18]. The notion of 'pathway activation' implied by the notion of pathway working units may thus violate intuition and hinder interpretation. Therefore we also extracted all pathway partitions which comply with the graph-theoretic notion of fully coupled flux [12]. It is known that the genes coupled by their enzymatic fluxes not only show similar expression patterns, but also share transcriptional regulators and frequently reside in the same operon in prokaryotes or similar eukaryotic multigene units such as the hematopoietic globin gene cluster. FCF is a special kind of network flux that corresponds to a pathway partition in which non-zero flux for one reaction implies a non-zero flux for the other reactions and vice versa. It is the strongest qualitative connectivity that can be identified in a network. The notion of an FCF is explained through an example in Fig. 4; for a detailed definition, see reference [12]. Again, a probeset falls in a list corresponding to a FCF if it is mapped to a KEGG node in some organism-ortholog of that FCF. To conclude, XGENE.ORG uses working units at various levels of generality. This hierarchy of units allows to capture and interpret biological issues that most strongly manifest in various kinds of existing biological models.

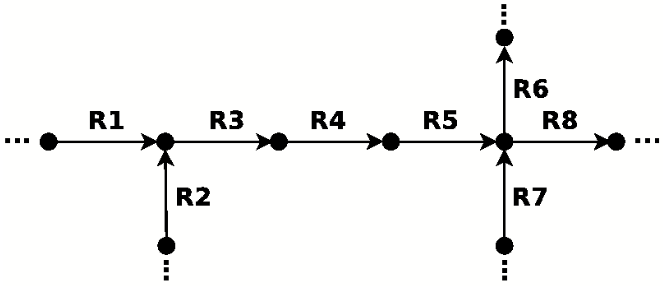


Figure 4. Fully coupled fluxes in a simplified network with nodes representing chemical compounds and arrows as symbols for chemical reactions among them. Each arrow can be labeled by a protein. R3, R4 and R5 are fully coupled as a flux in any of these reactions implies a flux in the rest of them. Note that R1 and R3 do not constitute a FCF as a flux in R3 does not imply a flux in R1.

The extraction of working units and computation of their activity in biological samples was conducted in Prolog. The process of computation of KEGG node activity in a sample set that originates from two different platforms is shown in Figure 5.

Currently, the aggregated activity of a unit in a sample is computed as the mean activity of all the measurement units that map on it in the given platform. When averaging is applied in Figure 5, it holds that $k_{wi} = (p_{xi} + p_{yi})/2 = g_{zi}$ and $k_{wj} = (p'_{aj} + p'_{bj} + p'_{cj})/3 = (g'_{dj} + 2g'_{ej})/3$. It means that the weight of gene g'_e is twofold with respect to g'_d as the former maps to two probesets while the latter to one probeset only. We are aware that averaging is an elementary aproach that may oversimplify the relationships and information transmission among units. Finding a biologically sound way to model the activity of genomic entities from microarray data is an open complex research issue. First of all, the mapping between probesets and genes is not unambiguous because the individual probesets map to more than one transcript dependent upon the biological condition [17]. There are efforts to refine the standard annotation of microarray probesets from gene level to transcript and protein level [22]. Secondly, it is advantageous to take into account internal structure of the modelled entities. More profound knowledge-based approaches to generalize towards more complex entities such as pathways can be found in [13, 15, 14]. However, these works always focus at a single type of applied knowledge and do not concern a universal workflow with multiple platforms on its input. Moreover, the application of such more sophisticated strategies to aggregate statistical values pertaining to subunits to represent analogical values of more general units is not scalable in the framework adopted by XGENE.ORG. In principle, this is because the simple average computation among subunits would have to be replaced by some sort of *subset selection*. Here, one searches for the best subset of subunits that best represent the parent unit, according to some optimality criterion. Generally, searching among subsets in a family of probesets S becomes quickly intractable with the growing size of S. For example, a selection of the best family of probesets for a given *gene* may be tractable as there is typically just a few probesets mapping to a gene in a platform. However, searching among subsets among in the pool of 10s-100s of probesets mapping to a *pathway* is generally no longer tractable.

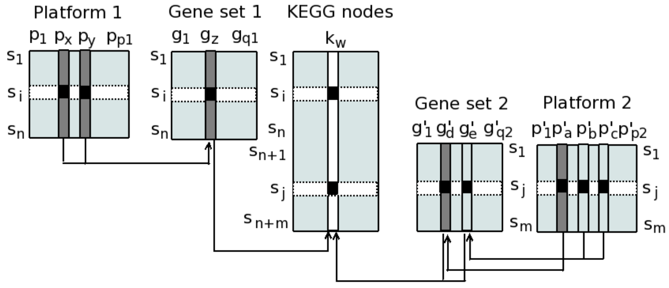


Figure 5. KEGG node activity. The activity of the node k_w in the sample s_i denoted as k_{wi} is given by the activity of its subordinate gene g_z whose activity is in turn given by the activity of its subordinate probesets p_x and p_y measured in Platform 1. The activity of the same node k_w in the sample s_j denoted as k_{wj} is given by the activity of g'_d and g'_e. The activity of g'_d is given by the activity of p'_a while activity of g'_e is inferred from activity of p'_b and p'_c measured in Platform 2.

3.2. Analysis Algorithms

After the collection of all data needed for a defined experiment, *normalization* is conducted separately for each involved platform to consolidate same-platform samples. Quantile normalization [2] ensures that the distribution of expression values across such samples is identical. As a second step, scaling provides means to consolidate the measurements across multi-platform samples. We subtract the sample mean from all sample components, and divide them by the standard deviation within the sample. As a result, all samples independently of the platform exhibit zero mean and unit variance. We conduct these steps using the Bioconductor [4] software.

After normalization, the most basic type of analysis that may be generated on user's request is *fold change* analysis whose goal is to rank the ability of the individual working units to distinguish among the user-defined classes. For this sake, we apply the one-way ANOVA (analysis of variance) method. In single platform tests where ANOVA ranks probesets, it determines if the sample distribution among classes has a significant effect on probe-set expression behavior. A significant p-value resulting from a one-way ANOVA test indicates that a probeset is differentially expressed in at least one of the classes analyzed. The lower p-value, the higher the probeset ranking. When ranking units of higher order, we do not proceed in a post hoc fashion from the single p-values of probe-sets but we model the expression of working units directly. For every unit, a complete list of probesets that map onto that unit is taken independently of the platform type. Their expression values in all the samples are gathered and factorized by the user-defined class variable[2]. With such prepared data, one-way ANOVA is run. Using the distinction for gene set statistical testing carried out in [5] we apply a self-contained test with subject sampling. No averaging is applied.

Having a single-tabular representation in which activity of a set of working units is computed across samples, a wide-scale of machine learning algorithms can be applied. The most interesting appear to be such algorithms that allow for direct human interpretation of the resulting models and still keep a good predictive power. Specifically, we included the J48 decision tree learner provided by the machine learning environment WEKA [21]. The K-nearest neighbor (kNN) algorithm from the same environment has also been included.

Finally, principal component analysis (PCA) is used for the purpose of dimensionality reduction in a space of working units with subsequent visualisation of samples [6]. PCA is known to retain those characteristics of the data set that contribute most to its variance. In XGENE.ORG it helps to exhibit class distribution in 2D and visually assess the potential of a set of working units to distinguish among classes.

4. Case studies

Here we demonstrate our methodology in two biological case studies. We address general tasks of tissue type classification. The first experiment focuses on distinct features of blood-forming (*hematopoietic*) and supportive (*stromal*) cellular compartments in the bone marrow. The second assesses differences in brain, liver and muscle tissues. Both experiments are of biological significance as they tackle novel challenges in understanding of cellular behavior: the former in the complex functional unit termed hematopoietic stem cell niche, where inter-dependent hematopoietic and stromal cell functions synergize in the blood-forming function of the bone marrow; the latter in comparison of cell fate determined by the tissue origin from the separate layers of the embryo: ectoderm (brain), endoderm (liver) and mesoderm (muscle). While of general character, the chosen tasks are not just random biological exercises as these studies may illuminate cellular functions determined by gene expression signatures in complex cell system seeded by cell-type-heterogeneous undifferentiated populations (hematopoietic and stromal stem cells in the cell niche), and in the cell-type-homogeneous differentiated tissues (brain, liver and muscle), respectively.

The significance tests at gene level identified elevated expression of genes canonical for the specific tissue studied, such as myelin basic protein in brain, isocitrate dehydrogenase in liver, tropomyosin in muscle and differential expression of integrin beta 5 inhematopoietic and stromal cell populations of the bone marrow.

The experiments with machine learning algorithms proved that working units applicable across platforms clearly distinguish among classes in both studies. The resulting models are compact, easy to interpret and accurate. Fig. 6 exemplifies the application of the decision tree learner J48 on the level of FCFs in the brain/liver/muscle study. The model tested by 10-fold cross-validation reaches the classification accuracy nearly 98%, it misclassifies 3 out of 131 samples. The tree has only 2 internal nodes (2 activity tests that put into use two FCFs) and 3 leaves (one leaf per class).

A similar conclusion follows from PCA visualizations (Fig. 7). The activity of working units tends to share the same pattern within classes as well as within the same platforms or the same laboratories. However, the class pattern is strong enough to clearly distinguish among classes independently of platform.

The complete overview of results is available via the XGENE.ORG webpage.

5. Discussion

XGENE.ORG is a web tool for analysis of gene expression data collected from heterogeneous (multi-platform) microarray platforms under the presence of genomic background knowledge. The integration of multi-platform data is conducted automatically by using the available genomic

[2] In Figure 5, the significance of k_w is inferred from concatenation of the expression vectors for probesets p_x, p_b, p'_a, p'_b and p'_c. The factorization is given by the sample distribution which is not shown.

background knowledge to define candidate working units general enough to be quantified in any sample regardless of the platform on which it was measured. The heterogeneous data are transformed into a single-tabular representation which summarizes the activity of the working units for all the collected samples. Such a unified representation lends itself to various types of analysis provided by XGENE.ORG based on statistical or machine learning methods.

The contribution of this tool is at least twofold. First, microarray experiments are costly, often resulting in numbers of samples insufficient for reliable modeling. The possibility of systematically integrating the experimenter's data with numerous public expression samples coming from heterogeneous platforms, would obviously help the experimenter. Second, such integrated analysis provides the principal means to discover biological markers shared by different-genome species.

Acknowledgments. The XGENE.ORG server code adopts parts of the open-source software Bioconductor [4] and WEKA [21]. For displaying pathway maps (such as in Fig. 7), XGENE.ORG uses data provided by the KEGG database [9]. The Czech-US coordination and travel of FZ is funded by the Czech Ministry of Education through project ME910. The authors are supported by the Czech Grant Agency through project 201/09/1665 (MH), the Czech Ministry of Education through projects MSM6840770038 (FZ) and MSM6840770012 (JK), and by the Children's Cancer Research Fund of the University of Minnesota (JT).

References

[1] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35(Database issue), January 2007.

[2] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003.

[3] T. G. O. Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25, 2000.

[4] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[5] J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[6] N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15):8409–8414, July 2000.

[7] D. W. Huang, B. T. Sherman, and R. A. Lempick. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4:44 – 57, 2009.

[8] D. W. W. Huang, B. T. T. Sherman, and R. A. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, November 2008.

[9] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32:277–280, 2004.

[10] A. Lewin and I. C. Grieve. Grouping gene ontology terms to improve the assessment of gene set enrichment in microarray data. *BMC Bioinformatics*, 7:426+, October 2006.

[11] V. Mootha, C. Lindgren, and S. L. et al. Pgc-1-alpha-responsive genes involved in oxidative phosphorylation are coorinately down regulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.

[12] R. A. Notebaart, B. Teusink, R. J. Siezen, and B. Papp. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLOS Computational Biology*, 4(1), 2008.

[13] J. Rahnenführer, F. S. Domingues, J. Maydt, and T. Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat Appl Genet Mol Biol*, 3, 2004.

[14] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35+, February 2007.

[15] G. Schramm, M. Zapatka, R. Eils, and R. König. Using gene expression data and network topology to detect substantial pathways, clusters and switches during oxygen deprivation of escherichia coli. *BMC Bioinformatics*, 8:149, 2007.

[16] A. Shaw and E. Filbert. Scaffold proteins and immune-cell signalling. *Nat Rev Immunol.*, 9(1):47–56, 2009.

[17] M. A. Stalteri and A. P. Harrison. Interpretation of multiple probe sets mapping to the same gene in affymetrix genechips. *BMC Bioinformatics*, 8:13+, January 2007.

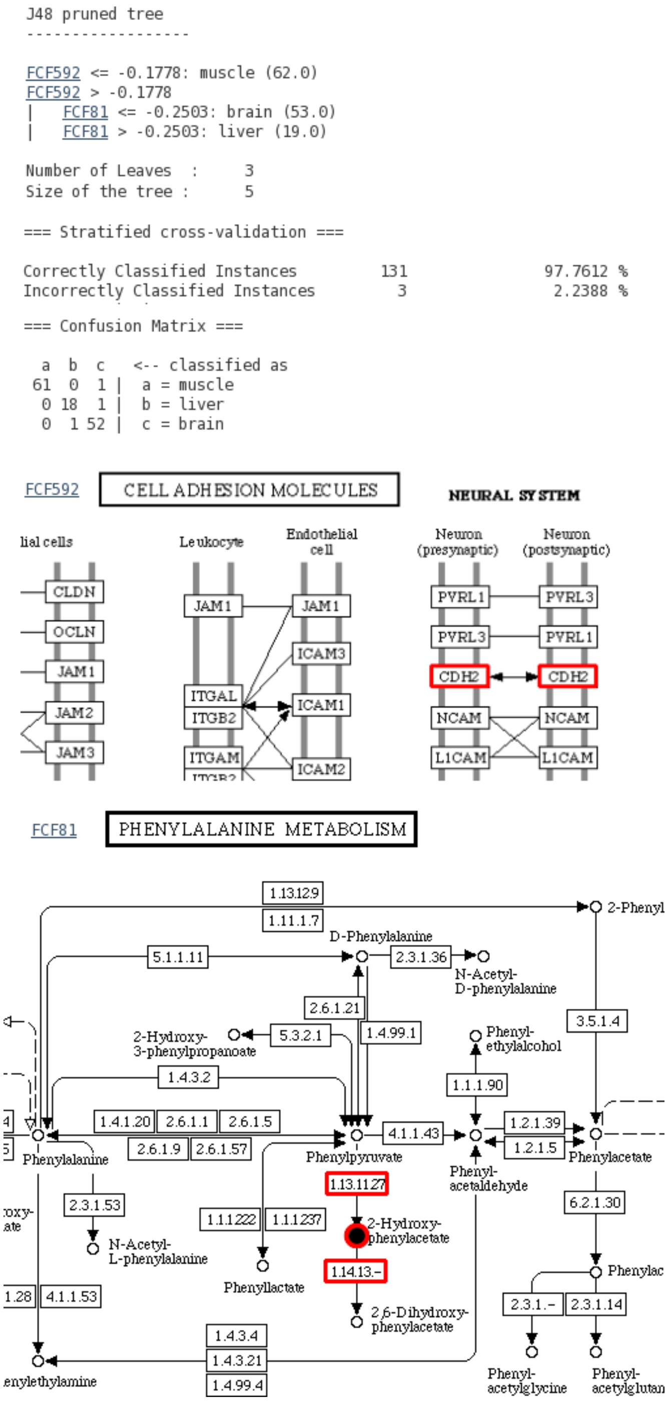
[18] Y. Y. Sun and J. Chen. mTOR signaling: PLD takes center stage. *Cell Cycle*, 7(20):3118–23, 2008.

[19] J. Tomfohr, J. Lu, and T. B. Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6, 2005.

[20] T. Weichhart and M. Semann. The PI3K/Akt/mTOR pathway in innate immune cells: emerging therapeutic applications. *Ann Rheum Dis.*, Suppl 3:iii:70–4, 2008.

[21] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.

[22] H. Yu, F. Wang, K. Tu, L. Xie, Y.-Y. Li, and Y.-X. Li. Transcript-level annotation of affymetrix probesets improves the interpretation of gene expression data. *BMC Bioinformatics*, 8:194+, June 2007.



```
J48 pruned tree
------------------

FCF592 <= -0.1778: muscle (62.0)
FCF592 > -0.1778
|   FCF81 <= -0.2503: brain (53.0)
|   FCF81 > -0.2503: liver (19.0)

Number of Leaves  :     3
Size of the tree :      5

=== Stratified cross-validation ===

Correctly Classified Instances         131               97.7612 %
Incorrectly Classified Instances         3                2.2388 %

=== Confusion Matrix ===

  a  b  c   <-- classified as
 61  0  1 |  a = muscle
  0 18  1 |  b = liver
  0  1 52 |  c = brain
```

Figure 6. The flux-based cross-platform decision tree for the brain/liver/muscle study. The tree is very compact, the class is determined by two activity thresholds on two fluxes, the fluxes are visualized using KEGG pathway maps (in bold).

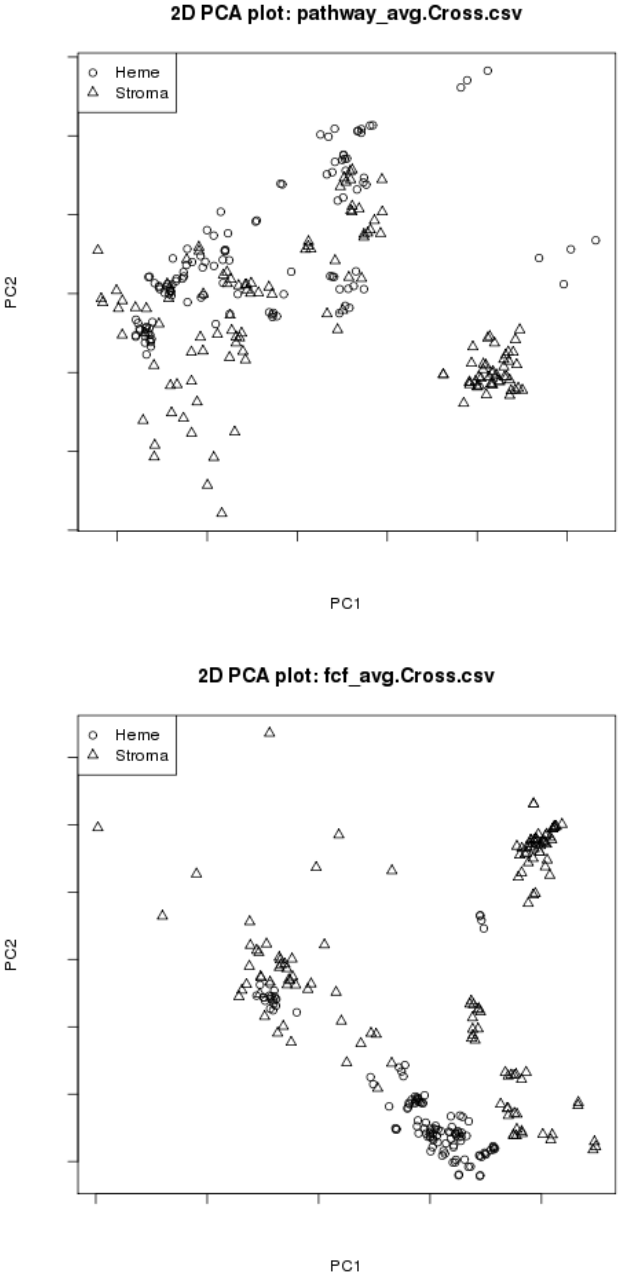


Figure 7. PCA in the hematopoietic/stromal study. The first subfigure shows cross-platform PCA in the space of pathways, the second subfigure uses FCFs instead. FCFs seem to better separate the classes (which is also confirmed by a higher classification accuracy if FCFs are used).