A Multi-strategy approach to informative gene identification from gene expression data

Liu, Ziying; Phan, Sieu; Famili, Fazel; Pan, Youlian; Lenferink, Anne E. G.; Cantin, Christiane; Collins, Catherine; O'Connor-McCourt, Maureen D.

Publisher’s version / Version de l'éditeur:
https://doi.org/10.1142/S0219720010004495
Journal of Bioinformatics and Computational Biology, 8, 1, pp. 19-38, 2010-02-01

Submitted to the Journal of Bioinformatics and Computational Biology

A MULTI-STRATEGY APPROACH TO INFORMATIVE GENE IDENTIFICATION FROM GENE EXPRESSION DATA

ZIYING LIU, SIEU PHAN, FAZEL FAMILI, YOULIAN PAN
Institute for Information Technology, National Research Council Canada, Ottawa, Ontario, K1A 0R6, Canada
{ziying.liu, sieu.phan, fazel.famili, youlian.pan}@nrc-cnrc.gc.ca

ANNE E.G. LENFERINK, CHRISTIANE CANTIN, CATHERINE COLLINS, MAUREEN D. O'CONNOR-MCCOURT
Biotechnology Research Institute, National Research Council Canada, Montreal, Quebec, H4P 2R2, Canada
{anne.lenferink, christiane.cantin, catherine.collins, maureen.oconnor}@cnrc-nrc.gc.ca

An unsupervised multi-strategy approach has been developed to identify informative genes from high throughput genomic data. Several statistical methods have been used in the field to identify differentially expressed genes. Since different methods generate different lists of genes, it is very challenging to determine the most reliable gene list and the appropriate method. This paper presents a multi-strategy method, in which a combination of several data analysis techniques are applied to a given dataset and a confidence measure is established to select genes from the gene lists generated by these techniques to form the *core* of our final selection. The remainder of the genes that form the *peripheral* region are subject to exclusion or inclusion into the final selection. This paper demonstrates this methodology through its application to an in-house cancer genomics dataset and a public dataset. The results indicate that our method provides more reliable list of genes, which are validated using biological knowledge, biological experiments and literature search. We further evaluated our multi-strategy method by consolidating two pairs of independent datasets, each pair is for the same disease, but generated by different labs using different platforms. The results showed that our method has produced far better results.

Keywords: Gene expression data analysis; Multi-strategy learning; Data mining and knowledge discovery.

1. Introduction

Discovering useful, and ideally, all previously unknown knowledge from historical or real-time data obtained from various sources, such as biological experiments or clinical information, is a complex and challenging task. This first requires an in-depth understanding of the domain and second the development of novel and appropriate strategies for data preprocessing and analysis. In high throughput genomics applications, knowledge discovery processes support various research and development activities. Two examples of these are: (*i*) the discovery of relationships between genes and their functions based on time-series data (such as a drug response over time or stages during embryonic development), and (*ii*) the investigation of gene responses to various treatments at one discrete time point. Many different statistic and data mining approaches have been developed and successfully applied to biological datasets.[1] However, a method suitable for analyzing one dataset may not be successful when used for another dataset. It is well recognized that the different methods for the identification of differentially expressed genes produce non-overlapping lists of genes. This has motivated many researchers to apply several techniques, instead of one. However several questions should be asked: (i) what are the suitable methods that can be applied, and (ii) how can the results generated by these methods be properly combined without losing any useful information. The objective of this paper is to address the latter question.

In this paper, we provide an overview of knowledge discovery in genomics with the emphasis on multi-strategy approaches in which a number of unsupervised learning techniques are applied to identify differentially expressed genes from a given dataset. The following sections consist of a brief summary of some related work followed by a detailed description of the biological problem which motivated us to consider a multi-strategy approach. We then describe our method, provide the results of its application on two cancer related datasets and additional evaluation that we performed on two pairs of independent datasets. We conclude that our multi-strategic method outperforms any individual method used in this study.

2. Related Work

The research reported in this paper is closely related to both supervised and unsupervised learning. This is in addition to bagging, boosting and randomization approach for the construction of ensembles of decision trees, which is well described in Dietterich's paper.[2] The topic of our paper also overlaps with feature selection based on multiple methods.

Supervised methods, which are mainly concept learners, generate hypotheses that are based on the original set of attributes. In many learning applications, the original learning space becomes inadequate. This inadequacy becomes evident through a high degree of irregularity in the distribution of instances and the models that are generated as output. Bloedorn *et al.* have developed a methodology to apply multiple learners and a range of strategies for an automated improvement of the knowledge representation space.[3] A system like AQ-17 has been able to significantly extend the machine learning capabilities as a multi-mechanism approach and produces a new generation of symbolic learning systems.[3] Hsu *et al.* proposed a high level optimization system (in the form of a wrapper) for relevance determination and constructive induction, and on integrating these wrappers with elicited knowledge on attribute relevance and synthesis.[4] Their approach is based on using decision support systems when multi-strategy machine learning is applied. Similarly, Geurts *et al.* proposed a new tree-based ensemble method for supervised classification and regression problems.[5] This approach consists of randomizing both attribute and cut-point choice while splitting a tree node. In the extreme case, this approach builds totally randomized trees whose structures are independent of the output values of the learning sample.

Similar efforts are seen in applying unsupervised methods for multi-strategy learning. Amershi and Conati outline a user modeling framework that uses both unsupervised and supervised machine learning in order to reduce development costs of building user models, and facilitate transferability.[6] They apply a framework to model student learning during interaction with the Adaptive Coach for Exploration (ACE) learning environment (using both interface and eye-tracking data). Learning from cluster examples (LCE) is a hybrid task for combining features of two common grouping tasks: learning from examples and clustering.[7] In this approach, each training example is a partition of objects. The objective is then to learn from a training set, a rule for partitioning unseen object sets. Multi-clustering is an example that has qualitative advantages over standard clustering when applied to vector-data images.

In biological applications, one of the most relevant publications regarding feature selection for biological datasets is the comparison and evaluation of ten different feature selection methods by Jeffery *et al*.[8] The authors applied the ten methods listed in their paper to nine microarray datasets and found that these methods returned very dissimilar gene lists; only 8-21% of the genes identified were common among the various methods. Most recently, Kadota *et al.* proposed a weighted average difference (WAD) method and compared it with seven other feature selection methods for ranking differentially expressed genes and found the strength of each method is dependent on the preprocessing method used prior to the feature selection.[9] Hua *et al.* also found that none of the feature selection methods they used are good across all scenarios.[10] Nevertheless, none of these papers tried to combine gene lists generated by various methods. Abruzzo *et al.* applied three feature selection methods to identify differentially expressed genes and combined genes generated by each method.[11] They concluded that a comprehensive list containing all of the differentially expressed genes can be obtained by applying multiple statistical tests and combing the results of these tests.

Random forest classification is another example of multi-strategy methods. Classifiers built using this approach consist of many decision trees where the end results are modes of the classes generated by individual trees. Diaz-Uriarte and Alvarez de Andres investigated the use of random forests for classification of nine microarray datasets.[12] They proposed a new method for gene selection based on random trees and bootstrapping that produced a relatively small set of genes while preserving predictive accuracy and concluded that this method is competitive with the existing methods.

3. The Multi-Strategy Approach

This section provides an overview of the proposed multi-strategy approach (Figure 1). Microarray data are first passed through a basic data preprocessing stage such as normalization, data filtering, and missing value handling. Certain procedures in data preprocessing require domain expertise and additional research. Data preprocessing helps us to better envision the scope of the knowledge discovery process and to ascertain whether the microarray experiments have been performed properly. The next step in the multi-strategy approach is to apply as many data analysis methods as desired to obtain the best possible lists of differentially expressed genes. The gene lists obtained from the selected methods are then consolidated based on a novel algorithm that we describe in this study.

The overall consolidation algorithm is summarized in Figure 2. After obtaining the gene lists from different analysis methods, we first established a confidence measure to select from these gene lists a set of genes to form the *core* of our final selection. The remainder genes form the *peripheral* set which is subject to exclusion or inclusion into the final selection by applying the method described below. Depending on the context of the problem under study, there is a variety of ways to define the confidence measure to form the *core* and *peripheral*. A simple confidence measure could be defined as a unanimity voting scheme, under which the *core* consists of genes that are identified by all methods that were applied. Another option is to define a less stringent voting scheme by defining the *core* as the genes that are selected by more than one method.

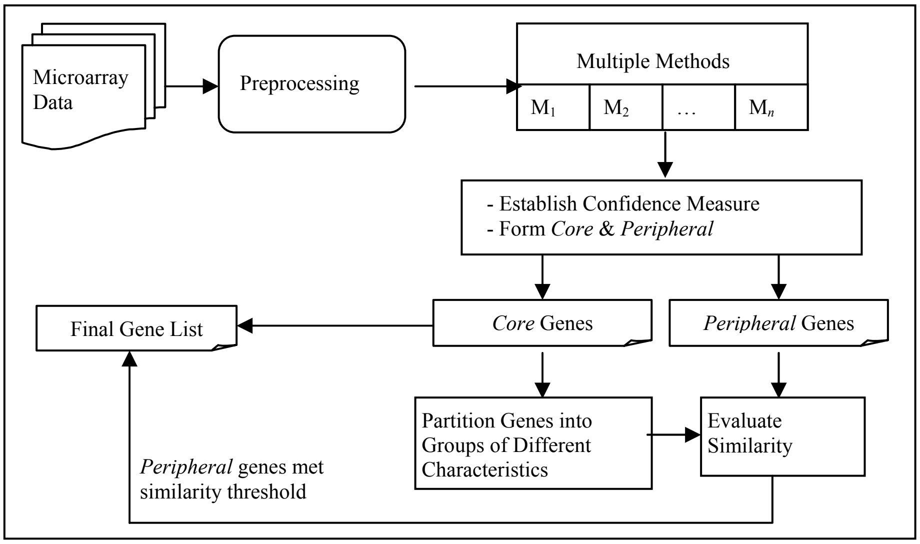
Microarray Data
Preprocessing
Multiple Methods
M_1
M_2
…
M_n
- Establish Confidence Measure
- Form *Core* & *Peripheral*
Final Gene List
Core Genes
Peripheral Genes
Partition Genes into Groups of Different Characteristics
Evaluate Similarity
Peripheral genes met similarity threshold

Fig. 1. The proposed multi-strategy approach for the analysis of microarray data.

The next step is the recruitment of similar genes from the *peripheral* into the *core*. This is done based on the principle of characteristic similarity such as gene co-regulation, pathways involved, and Gene Ontology (GO) annotation.[13] In order to achieve this, we first partition the genes in *core* into different characteristic groups. The genes in each group could:

i. participate in the same biological pathway (based on, for example, KEGG database),[14] or
ii. have the same biological function (based on GO annotation), or
iii. be regulated by the similar mechanism (based on common transcription factors).

We then evaluate the similarity between the genes in the *peripheral* and the characteristics of each group in the *core*. If a gene in the *peripheral* region passes the pre-established similarity threshold, this gene is recruited into the final gene list. During the functional clustering in the recruitment process, ideally, the *peripheral* genes and *core* genes are considered as disjoint sets. In practice, this is not always possible. In such case, we may consider the union (*core* + *peripheral*) and partition them into various functional clusters. If certain percentage of genes in a functional cluster shares their functions with a cluster in the *core* region, then the genes belonging to *peripheral* in the cluster are recruited, as indicated in our application below (Equation 1). The *core* genes and the recruited genes are derived from the differentially-expressed genes identified by the individual methods; they therefore inherit the statistical significance property from these original methods. If a gene in our final selected list is identified by more than one method, we could maintain a multiple significance measure for the gene; or if only one single measure is desired then for a conservative measure, we would assign the statistical significance for the gene to be the least significance among them.

```
1) Apply $n$ methods that generate $n$ lists of genes: $L_1, L_{2,} ...L_n$
2) Define confidence measure, $CM$
3) Form core, $L_{core}$, according to $CM$. For $CM$ defined as the less stringent majority voting
   model, we have
       $L_{core} = \bigcup_{i=1}^{n-1} \bigcup_{j=i+1}^{n} (L_i \bigcap L_j)$
4) Form peripheral, $L_{peripheral}$, according to $CM$:
       $L_{peripheral} = (\bigcup_{i=1}^{n} L_i) - L_{core}$
5) Partition $L_{core}$ into different characteristic groups $g = \{g_1, g_2, ...g_m\}$
6) Evaluate the similarity between genes in $L_{peripheral}$ and each group in $g$
7) Recruit genes from $L_{peripheral}$ into $L_{recruited}$ if the genes in $L_{peripheral}$ meet the
   similarity criteria
8) Build final gene set $L_{final} = L_{core} \bigcup L_{recruited}$
```

Fig. 2. Consolidation algorithm.

4. Data Used for This Study

We used two gene expression datasets for this study. The first dataset was obtained from a mouse mammary epithelial cell line that was treated with Transforming Growth Factor (TGF)-β,[15] whereas the second dataset is a leukemia dataset that was published by Stirewalt *et al.*[16] Each of these datasets has its unique properties and provides us with a good means to evaluate our proposed method.

4.1. *JM01 dataset*

This dataset was generated by exposing JM01 cell line to a treatment with the Transforming Growth Factor (TGF-β) for 24 hour.[15] TGF-β induces an Epithelial-to-Mesenchymal Transition (EMT) in these cells, a phenomenon characterized by significant morphology and motility changes, which are thought to be critical for tumor progression. TGF-β can act both as a tumor suppressor and tumor promoter depending on the context in which it is expressed. Given the opposite actions of TGF-β and the multiplicity of effectors that this growth factor utilizes, it is essential to identify those mediators that are specific to its tumor promoting or tumor suppressive pathways. Elucidation of the genetic programs underlying this EMT should provide a better understanding of the molecular mechanisms involved in cancer development and progression. The goal of this study is to identify the TGF-β modulated genes involved in the EMT process. The transcriptome changes after 24 hours in TGF-β treated vs. non-treated control cells were monitored using Agilent 41K mouse genome array (four technical replicates).

4.2. *Leukemia dataset*

This dataset contains the gene expression data of 38 healthy donors and 26 acute myeloid leukemia (AML) patients, and was obtained using the Affymetrix HG U133A microarray platform.[16] This analysis was focused on the identification of abnormally expressed genes in AML, which could serve as potential therapeutic targets. Seven over-expressed genes in the AML to normal (healthy donors) comparison were discovered in the original cohort. Quantitative RT/PCR studies of an independent set of normal and AML patient samples then confirmed the findings.[16] We also used these results for the validation of the outcome of our analysis.

5. Experimental Results

The effectiveness of the proposed methodology is demonstrated through its application to the two datasets (Section 4). After data preprocessing, we applied three most popular methods used in microarray data analysis: a parametric statistical method t-test, a variant of t-test called significant analysis of microarrays (SAM) and a non-parametric method Rank Products (RP) to identify lists of differentially expressed genes from these two datasets.[17-18] For t-test, the cutoff point is $p \leq 0.05$ for both datasets. When analyzing data using SAM, "one class response" was applied to JM01 dataset and "two class unpaired" to the Leukemia dataset. The FDR for SAM was 5% for both datasets and the analysis used 500 random permutations for both datasets. In our experiments with RP, the expected RP-values and False Discovery Rate (FDR) were calculated using 500 random experiments (number of permutations) of the same size of original datasets for both datasets. We selected genes based on 5% FDR for both datasets. To form the *core*, we selected genes that were identified by more than one method. The remaining genes that were identified by only one individual method fall into *peripheral*.

The DAVID annotation tool[19] was applied to partition the genes (Figure 3) in *core* (S4+S5+S6+S7) and *core* plus *peripheral* (S1+S2+S3+S4+S5+S6+S7), respectively based on similar characteristics (such as biological process, molecular function, sub-cellular location, protein-protein interaction, protein functional domains, pathway information, disease association and literature). This resulted in a set of *n* clusters, $\boldsymbol{g} = \{g_1, g_2, .., g_n\}$ for the genes in *core* and a set of *m* clusters, $\boldsymbol{G} = \{G_1, G_2, .., G_m\}$ for the genes in *core* plus *peripheral.* We applied $p \leq 0.05$ with each annotation in the enrichment analysis. For the clustering, with regards to KAPPA similarity threshold, we used 0.35 for JM01 dataset and 0.45 for the Leukemia dataset. The clusters in $\boldsymbol{g}$ and $\boldsymbol{G}$ were then pair-wise compared. The percentage of the genes in each cluster of $\boldsymbol{G}$ that are overlapped with each cluster of $\boldsymbol{g}$ was calculated as follows:

$$percentageOfOverlappedGenes(G_i, g_j) = \frac{numberOfGenes(G_i \cap g_j)}{numberOfGenes(G_i \cup g_j)} \times 100\% \quad (1)$$

If an overlap ≥ 70% is found for the pair (G_i, g_j), then all the *peripheral* genes in G_i are recruited into the final gene list. Although these thresholds are arbitrarily selected, we selected these thresholds in order to recruit reasonable number of genes.

We also applied promoter similarity measure for recruitment. For each individual gene in the *peripheral*, as long as it finds a gene in the *core* that has similar promoter composition with a similarity score ≥ 0.7, this gene is recruited. The promoter similarity was measured based on the number of the same transcription factor binding sites (TFBS) appeared in the promoter region of the compared pair of genes. Let χ and ψ be the sets of distinct TFBSs that are found from two promoter regions *X* and *Y*, respectively. Usually χ and ψ share some common instances. Let $n(Z)$ be the number of instances in a set *Z*, the similarity, *Sim(X, Y)*, between the two promoter regions *X* and *Y* can be defined as:

$$Sim(X, Y) = \frac{n(\chi \cap \psi)}{n(\chi \cup \psi)} \quad (2)$$

Since $\chi \cap \psi$ is a subset of $\chi \cup \psi$, $n(\chi \cap \psi) \leq n(\chi \cup \psi)$. The value of *Sim* is therefore between 0 and 1. We searched the significant ($p < 0.05$)[20] transcription factor binding sites using Profile Hidden Makov Model[21] based on positional weight matrices obtained from TRASFAC (release 2009.1)[22] from the promoter regions (1 kb upstream and 200 bp downstream of transcription start site).

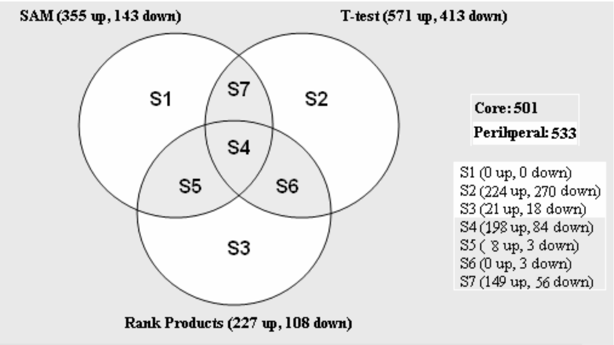

Fig. 3. Distribution of differentially expressed genes in the JM01 dataset identified by t-test, SAM and RP. (S4+S5+S6+S7) is the *core* and (S1+S2+S3) is the *peripheral*.

5.1 *JM01 dataset*

Figure 3 shows the distribution of differentially expressed genes identified by applying t-test, SAM and RP that make up the *core* (501 genes) and the *peripheral* (533 genes).

Table 1 shows the final number of up and down regulated genes identified by the multi-strategy approach. Initially, there were 355 up- and 146 down- regulated genes in the *core.* Through the recruitment processes, genes in the *peripheral* were brought in according to functional annotation clustering and promoter similarity. The final list contains 428 up- and 205 down- regulated genes. The recruitment is based on known (annotated) genes only.

Table 1. Final gene list for up and down regulated genes of JM01 dataset.

# of identified genes	TGF-β vs. Control	
	Up	Down
# of *core* genes	355	146
# of recruited genes	73	59
# of final genes	**428**	**205**

A Gene Set Enrichment Analysis (GSEA) was conducted as described in Subramanian's paper[23] to check the association of our final genes with human disorders/diseases or their related pathways. GSEA was performed using gene set databases C2 with 10,000 random gene set membership assignments. A summary of the results from GSEA is given in Table 2, where we list the gene sets with FDR < 0.25 and

Table 2. Summary of GSEA results of JM01 dataset.

Gene Set enriched in TGF-β treatment	Description	NOM p-val < 0.025	FDR < 0.25
BRENTANI_CELL_ ADHESION	Cancer related genes involved in cell adhesion and metalloproteinases	0.0012	0.08
TARTE_MATURE_PC	Genes overexpressed in polyclonal plastic cells (PPCs) as compared to mature plasma cells isolated from tonsils (TPCs) and mature plasma cells isolated from bone marrow (BMCs)	0.0064	0.21
CORDERO_KRAS_KD_ VS_CONTROL_UP	Genes upregulated in kras knockdown vs. control in a human cell line	0.0066	0.16
			FDR < 0.3
TGFBETA_ALL_UP	Upregulated by TGF-beta treatment of skin fibroblasts, at any timepoint	0.0246	0.30

nominal *p*-value < 0.025. To provide greater insight, we extended the analysis to include gene sets that met the FDR < 0.3. As can be noted, the results provides an overall picture of our final gene list, which is much more enriched in gene sets that are related to breast cancer, or TGF-beta signaling pathway, for example, BRENTANI_CELL_ADHESION, TGFBETA_ALL_UP. It also shows a consistency between our finding and the previous research.[23]

Table 3 shows a group of selected genes that were confirmed to be modulated by TGF-β. Several of these genes were independently confirmed by a proteomics analysis of TGF-β treated JM01 cells,[24] polymerase chain reaction (PCR), western blot or immunofluorescence microscopy. These results show that if only SAM was applied, eight (four up- and four down-regulated) genes would have been missed. The number of missed genes would have been two (one up- and one down-regulated) if only t-test was applied and thirteen genes (seven up- and six down-regulated) if only RP was applied.

Table 3. Biological validation by PCR, protemics approach and some others on certain selected genes of JM01 Data.

	Biological confirmation				TGF-β modulated, selected by different methods			
Gene Symbol	Proteomics	PCR	WB / IF microscopy	MRM spectroscopy	SAM	T_Test	RP	Multi-Strategy
Up-regulated genes								
Clu	x	x	x		x	x	x	x
Tnc		x			x	x	x	x
Ctla2a		x			x	x	x	x
Sdc3		x			x	x	x	x
Matn2		x			x	x	x	x
Fn1	x		x	x	x	x	x	x
Itga5	x		x		x	x	x	x
Acpp	x				x	x	x	x
Nes		x			x	x		x
Irf1		x			x	x		x
Fbln2				x	x	x		x
Col27a1		x			x	x		x
Pdgfra		x					x	x
Zyx		x				x		x
Snn		x				x		x
Itgb5	x		x			x		x
Down-regulated genes								
Tacstd2	x				x	x		x
Igfbp3		x			x	x		x
Cst6		x			x	x		x
Itga6	x	x	x				x	x
Cav1		x	x			x		x
Tk1		x				x		x
Ccne2		x				x		x

Notes: *PCR:* either semi-quantitative real-time PCR assays or real-time PCR assays; *WB*: western blot; *IF*: immunofluorescent microscopy; *MRM*: mass spectrometry-based multiple reaction monitoring.

These results clearly show that the application of a multi-strategy approach uncovers many important genes that could be missed when applying a single analysis method. For example, several members of the ITGB family have been implicated in the metastasis phenomenon,[25] whereas variations of caveolin-1 expression may have an important role in the progression of human breast lobular cancer,[26] and the c-kit and the PDGF receptors have been identified as potential targets for molecular therapy in breast cancer.[27]

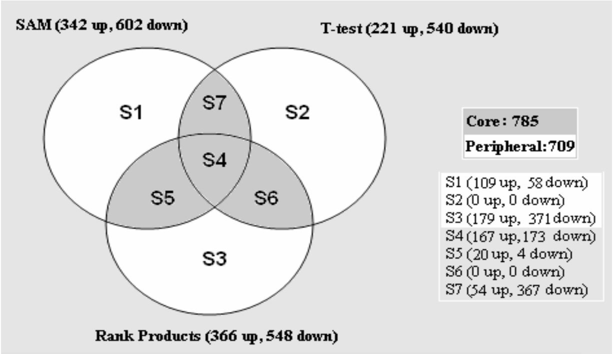

Fig. 4. Distribution of differentially expressed genes in the leukemia dataset identified by t-test, SAM and RP. (S4+S5+S6+S7) is the *core* and (S1+S2+S3) is the *peripheral.*

5.2. *Leukemia dataset*

Figure 4 shows the distribution of differentially expressed genes for the Leukemia dataset identified by applying t-test, SAM and RP, respectively. There are 785 and 709 genes in the *core* and the *peripheral,* respectively. The results of partitioning the *core* and the recruitment of additional genes from the *peripheral* are given in Table 4. Initially, there were 241 up- and 544 down-regulated genes in the *core*. Through the recruitment from the *peripheral*, 196 up- and 273 down- regulated genes were brought in according to functional annotation clustering and promoter similarity. The final list contains 437 up- and 817 down- regulated genes.

Table 4. Final gene list for up and down regulated genes of leukemia dataset.

# of identified genes	AML vs. Healthy donor	
	Up	Down
# of *core* genes	241	544
# of recruited genes	196	273
# of final genes	**437**	**817**

We then performed a GSEA using the same parameters for the cross checking of the final gene list as was done for the JM01 dataset. A summary of the results of this GSEA is given in Table 5, which lists gene sets with FDR < 0.01 and nominal p-value < 0.0001. These results show that the genes that are identified through the multi-strategy method are enriched, and that the list contains many gene sets that are strongly related to AML. This indicates that there is significant agreement between the gene list we generated using the multi-strategy method and previously published gene lists generated from leukemia datasets.[23] We therefore strongly believe that our multi-strategy approach for selecting genes is more reliable.

We validated the results of multi-strategy method using two leukemia databases. The first database is LeGenD,[28] a leukemia database curated from several published biomedical research literature databases (PubMed, OMIM, CGAP etc). The second database is GeneCards,[29] which is curated according to the information from biomedical research literature databases (*OMIM, SWISS-PROT, Genatlas, GeneTests, GAD, GDPInfo, bioalma, Leiden, Atlas, BCGD, TGDB and/or HGMD*). Table 6 shows the result of this analysis and indicates that several of the genes selected through our multi-strategy method are actually already identified leukemia disease genes. These results also show that if only SAM was applied, five genes would have been missed, whereas seven would have been missed if only t-test was applied and two would have been missed if only RP was applied. The multi-strategy approach therefore helped us to identify many important genes that would have otherwise been missed. T-cell leukemia/lymphoma protein (TLX1), for example, has been shown to be involved in a chromosomal aberration that may be a cause of the acute T-cell lymphoblastic leukemia (T-ALL).[30] In addition, a chromosomal aberration involving ARHGEF12 may be the cause of acute leukemia.[30] For example, the translocation of t(11;11)(q23;23) with the mixed-lineage (MLL) gene is observed to form myeloid/lymphoid fusion partner in acute myeloid leukemia.[30] Lastly, translocation of FOXO3 gene with the MLL gene is associated with secondary acute leukemia.[30] Table 6 also shows that seven of the identified genes through our multi-strategy approach were confirmed by qRT-PCR studies of an independent set of normal and AML patient samples.[16] The GSEA results have provided a clear insight in the significance of genes identified through the multi-strategy approach. The validation of these results using the two leukemia databases provided solid confirmation. Overall, these findings strongly suggest that multi-strategy approach successfully identify differentially expressed genes with a higher degree of reliability.

Table 5. Summary of GSEA results of Leukemia dataset.

Gene Set enriched in AML	Description	NOM p-val < 0.001	FDR < 0.01
ROSS_MLL_FUSION	Genes that distinguish pediatric acute myeloid leukemia subtypes with MLL chimeric fusion genes	< 0.0001	0.0000
ALCALAY_AML_ NPMC_UP	Increased expression in NPMc+ leukemias	< 0.0001	0.0003
ROSS_PML_RAR	Genes that distinguish pediatric acute myeloid leukemia subtypes t(15;17)[PML RAR-alpha].	< 0.0001	0.0003
IGLESIAS_E2FMINUS_UP	Genes that increase in the absence of E2F1 and E2F2.	< 0.0001	0.0094
			FDR < 0.025
VERHAAK_AML_NPM1_ MUT_VS_WT_UP	Genes that are upregulated in AML NPM1 mutant versus AML NPM1 wild type	< 0.0001	0.0242

Table 6. Biological validation by PCR, leukemia databases on selected genes of leukemia data.

	Biological confirmation			Important Leukemia genes selected by different methods			
Gene Symbol	Q-RT/PCR*	LeGenD	GeneCard	SAM	T_Test	RP	multi-strategy
Up-modelated genes							
BAX		x	x	x	x	x	x
CAV1		x		x	x	x	x
CEBPA		x	x	x	x	x	x
FLT3			x	x	x	x	x
PBX3		x		x	x	x	x
RUNX1		x	x	x	x	x	x
NQO1			x	x	x		x
TLX1		x	x	x			x
HOMER3	x			x	x	x	x
BIK	x			x	x	x	x
CCNA1	x			x	x	x	x
WT1	x			x	x	x	x
IL3RA	x			x	x	x	x
JAG1	x			x	x	x	x
FUT4	x			x	x	x	x
Down-modulated genes							
CLC		x		x	x	x	x
GATA1			x	x	x	x	x
MLLT3		x		x	x	x	x
MYC		x		x	x	x	x
PBX1		x	x	x	x	x	x
RGS2		x		x	x	x	x
STAT5B			x	x	x	x	x
TAL1		x	x	x	x	x	x
HLF			x	x	x	x	x
IRF1		x	x	x	x	x	x
NOTCH1		x	x	x	x	x	x
RARA			x	x	x	x	x
SETBP1			x	x	x	x	x
TCL1A			x	x	x	x	x
ARHGAP26			x	x		x	x
ARHGEF12		x	x			x	x
CDC23		x				x	x
FOXO3		x				x	x
MLL			x			x	x
MSH2			x			x	x

Note: Q- RT/PCR validation in Stirewalt's study.[16]

6. Additional evaluations

To further demonstrate the effectiveness of the proposed methodology, we performed additional experiments on two pairs of independent datasets using Jaccard coefficient. Each pair of datasets deals with the same disease phenotypes, but produced by different labs using different platforms. The first pair is about lung cancer, consisting of two data sets, one with 13 squamous cell lung cancers and 5 normal lung specimens[31] and the other with 21 squamous cell lung cancers, and 17 normal lung specimens[32]. The second pair is about Duchenne muscular dystrophy (DMD) consisting of two data sets, one with 12 DMD patient samples and 12 unaffected control[33] and the other with 22 DMD patient samples and 14 control samples[34]. During data pre-processing, the cDNA data was log2 transformed while quantile normalization was applied to all datasets, except the Affymetrix GeneChip data, that was pre-processed using RMA[35]. We then applied RP, SAM, T-test and our Multi-Strategy method to identify the significant genes in each dataset. In each pair, the genes present in both datasets were used in the subsequent analysis.

Zhang *et al.* proposed a *POGR* (percentage of overlapped genes related) score[36] to evaluate the reproducibility as $POGR_{12} = (k+O_{r12})/\ l_1$ or $POGR_{21} = (k+O_{r21})/\ l_2$, where $k = l_1 \cap l_2$ represents the common genes between lists l_1 and l_2, O_{r12} represents the number of genes in list l_1 that are not in k but are functionally related with at least one gene in list l_2. Similarly, O_{r21} represents the number of genes in list l_2 that are not in k but are functionally related with at least one gene in list l_1. We extended their approach in Jaccard coefficient as:

$$J_{POGR} = (k + O_{r12} + O_{r21})/(l_1 \cup l_2) \tag{3}$$

This way, not only we count the number of genes overlapped between the two lists based on their gene expression, but also we consider functionally related genes. The functionally related genes were identified using DAVID functional classification tool[19] with similarity 0.35. We used Matchminer software[37] to process the gene identifiers from different datasets/platforms. Then we calculated the Jaccard coefficient (Equation 3) for each pair of gene lists generated by all the four methods (Table 7). The Jaccard coefficient is the highest when using multi-strategy method for both pair of datasets. Therefore, by recruiting functionally related genes, the multi-strategy method generates the two lists that are in much better agreement than any individual method used in this study.

Table 7. Jaccard coefficient of the each method.

	Lung Cancer	DMD
T-Test	31.76%	36.06%
SAM	27.13%	37.11%
RPs	63.79%	81.07%
Multi-strategy	**69.06%**	**87.95%**

For Lung Cancer datasets, t-test, p<=0.01, SAM and RPs with FDR1%, For DMD datasets t-test, p<=0.001, SAM and RPs with FDR 0.1%,

7. Discussion and Conclusions

In the past decade, many statistical methods have been developed to identify differential expressed genes during embryonic development, disease vs. normal, etc. Since the last few years, researchers have come to realize that application of a single analysis method may not identify all genes that are involved in the process that is studied. We therefore propose a multi-strategy analysis method which integrates the output of several analysis methods, thereby generating a more comprehensive result. In this paper, we have applied a multi-strategy method to the two microarray datasets. The results of our data analysis and further evaluation based on two pairs of independent datasets have shown that our consolidation method performed better than any single participating method; it is able to identify additional important genes that would be missed when applying only one method. This method provides a stronger confidence in the results without incurring high false identification.

The main challenge in all multi-strategy methods is the consolidation and interpretation of the results. We strongly believe that given the objectives of this research and the biological problem, applying multiple analysis methods has produced better results than any single method. The two most important questions arising from the analysis we described here are: (i) what are the most appropriate methods that one needs to apply to search for list of informative genes that would lead to identifying biomarkers, and (ii) how to combine the generated results. We are pursuing this research further by applying this approach to other datasets, and are also evaluating other comparison methods. We believe that the novelty of our approach is in employing data mining concepts, refining them and combining them with other existing methods in the field of unsupervised learning to discover useful and more comprehensive information from biological data.

One has to keep in mind that the *core* of our approach is an *unsupervised* method where: (*i*) the data is not labeled, (*ii*) no confusion matrix can be applied, and (*iii*) the real decision on what is False Positive and what is False Negative can be made through: (*a*) domain expert feedback, (*b*) literature validation, (*c*) biological experimental confirmation, and (*d*) perhaps follow up studies, such as performing additional biological experiments, etc. Because of the cost associated with biological experiments, we were not able to validate all the identified genes experimentally. We have done our best to identify which genes, from amongst all the genes that we identified to be the most informative, are truly the most relevant (or true positives) based on experimental confirmation (see Table 3).

8. Supplementary Materials

Gene lists are available upon request to the corresponding author. The gene lists include detailed list of up-down *core* gene lists and recruited gene lists by different methods.

9. Acknowledgement

We would like to thank Miroslava Cuperlovic-Culf and two anonymous reviewers for the constructive comments. This research is partially supported by Genomics and Health Initiative of the National Research Council Canada.

References

1. Cuperlovic-Culf M, Belacel N, Ouellette R, Determination of Tumour Marker Genes from Gene Expression Data, *Drug Discovery Today* **10** (6): 429-437, 2005.

2. Dietterich TG, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, *Machine Learning* **40**(2): 1-19, 2000.
3. Bloedorn E, Michalski RS, Wnek J, Multistrategy constructive Induction: AQ17-MCI, *Proceedings of the Second International Workshop on Multistrategy Learning (MSL93)*, Harpers Ferry, pp.188-203, Morgan Kaufmann, 1993.
4. Hsu WH, Welge M, Redman T, Clutter D, High performance commercial data mining: A multi-strategy machine learning application, *Data Mining and Knowledge Discovery* **6**(4): 361-391, 2002.
5. Geurts P, Ernst D, Wehenkel L, Extremely randomized trees Source**,** *Machine Learning* **63**(1): 3-42, 2006.
6. Amershi S, Conati C, Unsupervised and supervised machine learning in user modeling for intelligent learning environments, *Proceedings of the 12th international conference on Intelligent user interfaces IUI '07*, ACM, pp. 72-81, 2007.
7. Kamishima T, Motoyoshi F, Learning from Cluster Examples, *Machine Learning,* **53**(3): 199-233, 2003.
8. Jeffery IB, Higgins DG, Culhane AC, Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data, *BMC Bioinformatics* **7**:359, 2006.
9. Kadota K, Nakai J, Shimizu K, A weighted average difference method for detecting differentially expressed genes from microarray data, *Algorithms for Molecular Biology* **3**:8, 2008.
10. Hua J, Tembe WD, Doughertya ER, Performance of feature-selection methods in the classification of high-dimension data, *Pattern Recognition* **42**(3): 409 – 424, 2009.
11. Abruzzo LV, Wang J, Kapoor M, Medeiros LJ, Keating MJ, Highsmith WE, Barron LL, Cromwell CC, and Coombes KR, Biological Validation of Differentially Expressed Genes in Chronic Lymphocytic Leukemia Identified by Applying Multiple Statistical Methods to Oligonucleotide Microarrays, *J Mol Diagn* **7**(3): 337–345, 2005.
12. Diaz-Uriarte R, Alvarez de Andres S, Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* **7**:3, 2006.
13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, The Gene Ontology Consortium: Gene ontology: tool for the unification of biology, *Nature Genetics* **25**(1): 25-29, 2000.
14. Kanehisa M, Goto S, KEGG: Kyoto Encyclopedia of Genes and genomes, *Nucleic Acids Res* **28**(1):27-30, 2000.
15. Lenferink AEG, Magoon J, Cantin C, O'Connor-McCourt MD, Investigation of three new mouse mammary tumor cell lines as models for transforming growth factor (TGF)-β and Neu pathway signaling studies: identification of a novel model for TGF-β-induced epithelial-to-mesenchymal transition, *Breast Cancer Res* **6**(5): 514–530, 2004.
16. Stirewalt D, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, Cronk MR, Dorcy KS, McQuary MR, Hockenbery D, Wood D, Heimfeld S, Radich JP, Identification of genes with abnormal expression changes in acute myeloid leukemia, *Genes Chromosomes Cancer* **47**(1): 8-20, 2008.
17. Tusher VG, Tibshirani R, Chu G, Significance analysis of microarrays applied to the ionizing radiation response, *PNAS* **98** (9): 5116-5121, 2001.

18. Breitling R, Armengaud P, Amtmann A, Herzyk P, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *FEBS letters* **573**(1-3): 83-92 2004.
19. Jr GD, Sherman BT, Hosack DA, Yang Y, Gao W, Lane HC, Lempicki RA, Software DAVID: Database for Annotation, Visualization, and Integrated Discovery, *Genome Biology* **4**(5): R60, 2003.
20. Pan Y, Phan S, Threshold for positional weight matrix, *Engineering Letters* **16**(4): 498-504, 2008.
21. Eddy SR, Profile hidden Markov models, *Bioinformatics* **14**(9): 755-763, 1998.
22. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE and Wingender E, TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res* **34**: 108-10, 2006.
23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *PNAS* **102**(43): 15545–15550, 2005.
24. Hill JJ, Tremblay TL, Cantin C, O'Connor-McCourt MD, Kelly JF, Lenferink AEG, Glycoproteomic analysis of two mouse mammary cell lines during transforming growth factor (TGF)-β induced epithelial to mesenchymal transition, *Proteome Science* **7:**2 2009.
25. Pontier SM, Muller WJ, Integrins in breast cancer dormancy, *APMIS* **116**(7-8): 677-684 2008.
26. Perrone G, Altomare V, Zagami M, Morini S, Petitti T, Battista C, Muda AO, Rabitti C, Caveolin-1 expression in human breast lobular cancer progression, *Mod Pathol* **22**(1):71-8 2009.
27. Roussidis AE, Theocharis AD, Tzanakakis GN, Karamanos NK, The importance of c-kit and PDGF receptors as potential targets for molecular therapy in breast cancer, *Current medicinal chemistry* **14**(7): 735-743 2007
28. Bioinformatics Organization, Inc. http://www.bioinformatics.org/legend/legend.htm. Accessed on March, 2009.
29. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D, GeneCards: integrating information about genes, proteins and diseases, *Trends Genet* **13**(4): 163 1997.
30. The UniProt Consortium, The Universal Protein Resource (UniProt), *Nucleic Acids Res* **35**(Database issue): D193-7 2007
31. Garber ME, Troyanskaya OG, Schluens K, *et al.* Diversity of gene expression in adenocarcinoma of the lung, *Proc Natl Acad Sci USA* **98**:13784–13789 2001
32. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proc. Natl. Acad Sci. USA.* **98**:13790–13795. 2001
33. Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, and Kunkel LM. Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle, *Proc. Natl. Acad Sci. USA.* **99**:15000–15005 2002
34. Pescatori M, Broccolini A, Minetti C, Bertini E, Bruno C, D'Amico A, Bernardini C, Mirabella M, Silvestri G, Giglio V, Modoni A, Pedemonte M, Tasca G, Galluzzi G, Mercuri E, Tonali PA, Ricci E. Gene expression profiling in the early phases of DMD: a constant

molecular signature characterizes DMD muscle from early postnatal life throughout disease progression, *FASEB J.* **21**,1210-1226 2007

35. Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed. Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research* **31**(4):e15 2003
36. Zhang M, Zhang L, Zou JF, Yao C, Xiao H, Liu Q, Wang J, Wang D, Wang CG and Guo Z, Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes, *Bioinformatics* **25**(13):1662-1668, 2009
37. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, Zeeberg B, Ajay and Weinstein JN, MatchMiner: a tool for batch navigation among gene and gene product identifiers, *Genome Biology*, **4**(4):R27, 2003

Ziying Liu received a Master's degree in medical sciences in Katholieke Universiteit Leuven, Belgium in 1994 and a Master's degree of System Science at the University of Ottawa in 2006. She is currently working in the Knowledge Discovery Group at the Institute for Information Technology, National Research Council Canada. She is interested in data mining, knowledge discovery, and biomedical database applications.

Dr. Sieu Phan is a senior research scientist at the National Research Council Canada. His research interests include artificial intelligence, data mining, knowledge discovery, communication network management, and intelligent decision-support systems. He is currently a co-lead of the Cancer Omics Project, at the Knowledge Discovery Group, Institute for Information Technology, focusing on identifying omics signatures that could be used to predict cancer risk and therapeutic responses.

Fazel Famili is a Group Leader for the Knowledge Discovery group, working at the Institute for Information Technology (IIT) of the National Research Council of Canada, where he has been working for the past 24 years. Dr. Famili has been actively involved in the field of Artificial Intelligence, Data Mining and Bioinformatics and successful applications of these technologies. He has a strong data mining and bioinformatics team within IIT (>22 staff) that is currently engaged in unique research and development in data mining for genomics, proteomics and health care. His research has been on data mining, machine learning and bioinformatics and their applications to real world problems in various data rich environments, such as life sciences. Dr. Famili has edited two books, has published over 50 articles in the area of data mining and AI and has a US/Canadian data mining patent. He is also an adjunct professor at the University of Ottawa.

Youlian Pan holds a Ph.D. in Biology and a Master in Computer Sciences, both from Dalhousie University, Halifax, Canada. He is currently a research officer at National Research Council Canada. His research interests include data mining, machine learning and knowledge discovery. His specific interests are in the integrative data mining approaches with biological and medical applications.

Anne EG Lenferink obtained her Ph.D. in Cell Biology at the Katholieke Universiteit Nijmegen (Radboud Universiteit Nijmegen) and is currently a Research Officer in the Receptors, Signaling and Proteomics group at the Biotechnology Research Institute (BRI), National Research Council of Canada in Montréal, Québec. Her research interests include structure-function relationships of proteins, proteomics, cancer cell biology, bioinformatics and the design and execution of animal studies for evaluation of novel anti-cancer drugs.

Christiane Cantin has been a Technical Officer at the NRC-BRI since 1997 where she specializes in new protocol development and optimization, especially in mammalian cell biology, microarrays and immunohistochemistry.

Catherine Collins is a Technical Officer working in the Receptors, Signaling and Proteomics group of the Biotechnology Research Institute, National Research Council of Canada in Montréal, Québec.

Dr. Maureen O'Connor-McCourt has been a Senior Research Officer and Group Leader of the Receptor, Signaling and Proteomics group at the BRI since 1996. She has been a member of Protein Engineering Network Centres of Excellence (PENCE) Montréal, Canada, since 1990, as the Chairman from 1994 to 1996 and was a member of the Training and Education Committees from 1990 to 1994. Her research focuses on the identification and study of protein-protein interactions in growth-mediated tumorigenesis, development of anti-cancer drugs and therapy, as well as imaging and non-imaging diagnostics. Her laboratory has patented the software program SPRevolution, crucial for biosensor analysis. Dr. O'Connor-McCourt has published over 60 scientific papers. She also reviews manuscripts for a number of journals and evaluates grant applications for the CIHR and FRSQ, and is an external thesis committee member for the University of Toronto, the University of Ottawa and McGill University.