Robust PCA improves biomarker discovery in colon cancer with incorporation of literature information

Thomas Minta and De Moor Bart*[1]

[1]ESAT-SCD / iMinds Future Health Department Katholieke Universiteit Leuven,3001, Leuven,Belgium.

Email: minta.thomas@esat.kuleuven.be;

*Corresponding author

Abstract

Microarray technology, which can handle thousands of genes of several hundreds of patients at a time, makes it hard for scientists to manually extract relevant information about genes and diseases, especially cancer. Bioinformatics and statistical methods collect the appropriate information from these data sets in light of the specific application. To date, many research groups identified differentially expressed genes based on microarray data and literature information. Still, the lack of efficient methods for assessing the biological implications of gene expression data remains an important difficulty in exploiting these information.

In this study, we propose a data integration approach in which microarray expressions values are weighted with literature information. Initially we present generalized eigenvalue decomposition(GEVD) in terms of ordinary eigenvalue decomposition (EVD) for the integration of microarray and literature information. Then robust PCA (RPCA) was used for the identification of differentially expressed genes of colon cancer. Finally we search for genes which are overlapped in co-expressed genes of differentially expressed ones and colon cancer data.

In colon cancer, the highest leave-one-out (LOO) areas under the receiver operating characteristic curves (AUC) were obtained with differentially expressed genes and co-expressed genes together on prediction ranged from 0.868 to 0.989.

In this case, the prediction of outcomes improved when literature information was considered. This suggests that incorporation of external knowledge into the analysis of microarray improves the identification of cancer specific genes. This emphasizes the need for comprehensive multi-modal data in gene expression analysis.

1 Introduction

Biomarker discovery and prognosis prediction are essential for improved personalized treatment of cancer. Microarray technology is a significant tool in gene expression analysis and cancer diagnosis. It can simultaneously handle thousands of genes. Microarray data are typically used for class discovery [1], [2] and prediction [3], [4]. The challenge in dealing with microarray data lies in the fact that there is a difference in orders of magnitude between the number of samples (typically less than hundred) and the number of genes (typically tens of thousands). The measurements also contain both measurement and systematic noise with an impact on classification accuracy.

An ever increasing number of techniques are available for the discovery of clusters of samples with similar gene expressions in microarray data. Troyanskaya and colleagues developed nonparametric methods to identify differentially expressed genes (DEG) in microarray data [5]. Principal component analysis (PCA) and PCA-based approaches were used as well for the identification of DEG in pulmonary adenocarcinoma [6] and E coli [7]. Besides well-known statistical tests such as the chi-square test [8], Chun and Colleagues proposed a new test, the 'half Student's t-test', specifically for detecting DEG in heterogeneous diseases [9]. Generalized Eigen Value Decomposition has been shown to have great potential within bioinformatics for extracting common information from datasets such as genomics and proteomics data [10] [11]. Recently robust PCA (RPCA) based method for discovering differentially expressed genes was proposed in [12]. Several studies have used microarray and literature information together to obtain differentially expressed genes with elements for biological understanding and for generating and validating new biological hypothesis [13–15].

In this paper, we propose a data integration strategy in which gene expression values are weighted with literature information. We represent generalized eigenvalue decomposition (GEVD) in terms of eigenvalue decomposition (EVD) for the integration of two data sets. Then we apply robust PCA (RPCA) to discover the differentially expressed genes in colon cancer. Finally we assume that there are disease-specific genes that were not identified in the analysis due to lack of literature information. Thus we search GeneMANIA [16] for obtaining co-expressed genes related to disease specific genes.

2 Data sets

2.1 Colon Dataset

The colon cancer data set investigated in this manuscript was taken from the Bioinformatics Research Group Repository [17]. It contains 62 samples, among them 40 colon tumor samples and 22 normal colon samples analyzed with an Affymetrix Oligonucleotide array complementary to more than 6,500 human genes [18]. Here we work with 1,988 genes and 12 controls of highest minimal expression over the samples. Data were standardized to

a mean of 0 and standard deviation of 1.

On classification, 2/3rd of the data samples of each class are assigned randomly to the training and the rest to the test set. These randomization are the same for all numerical experiments on all data sets. This split was performed stratified to ensure that the relative proportion of outcomes sampled in both training and test set was similar to the original proportion in the full data set. In all these cases, the data were standardized to zero mean and unit variance.

2.2 Literature Information

For colon cancer, a well defined cancer vocabulary was created, with 2406 terms from NCI Dictionary of Cancer Terms [19]. Pubmed abstracts with the terms in the vocabulary were extracted using Perl, version 5. 10. 1 for windows. Literature Information was created with rows corresponding to genes, and columns to cancer related terms. We chose to retrieve entries containing the official gene name, abbreviations or aliases in the the corresponding field. Each entry in the matrix, corresponds to the number of Pubmed abstracts in which the gene and term co-occur. For each gene, the result of query is downloaded in HTML format. All Pubmed abstracts linked to these genes are then indexed using the cancer vocabulary. Gene vectors are then constructed by averaging the vectors of all the abstracts associated to that gene. Finally, the cosine similarity measure is used to obtain gene-to-gene distances between 0 and 1.

3 Method

3.1 Generalized Eigenvalue Decomposition

The generalized singular value decomposition (GSVD) of $m \times N$ matrix A and $p \times N$ matrix B is [20] is obtained from the singular value decomposition (SVD) of matrix $[A; B]$

$$A = U\Sigma_A X^T \tag{1}$$

$$B = V\Sigma_B X^T \tag{2}$$

where U, V are orthogonal matrices and columns of X are generalized singular vectors.

If $B^T B$ is invertible, then the GEVD of $A^T A$ and $B^T B$ can be obtained from Equations (1) and (2) as follows:

$$A^T A(X^T)^{-1} = B^T B(X^T)^{-1}\Lambda. \tag{3}$$

where Λ is a diagonal matrix with diagonal entries $\Lambda_{ii} = (\frac{\Sigma_{A_{ii}}}{\Sigma_{B_{ii}}})^2$, $i = 1, \ldots, N$.

Equation (3) can be represented in eigenvalue decomposition (EVD) as follows:

$$(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}U = U\Lambda.$$

where $U = (B^TB)^{1/2}(X^T)^{-1}$. Let $D = A(B^TB)^{-1/2}$, then SVD of matrix D is given below:

$$D = V\Lambda U^T.$$

The principal component analysis (PCA) [21] of $m \times n$ matrix D equals

$$D = TC^T,$$

where score $T = V\Lambda$, is a $m \times n$ matrix and coefficients $C = U$, is an orthogonal $n \times n$ matrix.

The matrix $(B^TB)^{-1/2}$ is defined [22] as follows: Let EVD of $B^TB = T\Sigma T^T$, where columns of T are eigenvectors and Σ is a diagonal matrix. $(B^TB)^{1/2} = T\Sigma^{1/2}T^T$ and $(B^TB)^{-1/2} = TQT^T$, where Q is a diagonal matrix with diagonal entries $Q_{ii} = (\Sigma_{ii})^{-1/2}$, $i = 1, \ldots, N$.

3.2 Robust Principal Component Analysis (RPCA)

Robust principal component analysis (RPCA) proposed by Candes and colleagues [23]. Let $\|C\|_* = \Sigma_i\sigma_i(C)$ denote the nuclear norm of the matrix C, that is, the sum of its singular values, and let $\|S\|_1 = \Sigma_{ij}|S_{ij}|$ denote l_1-norm of S. Supposing that D denotes the observation matrix, RPCA solves the following optimization problem:

minimize $\|C\|_* + \lambda\|S\|_1$, subject to $D = C + S$

where λ is a positive regulation parameter. For the above equation for RPCA problem, Lin and colleagues gave a method for solving it, which is referred to as the inexact ALM(IALM) method [24]. The details of this algorithm can be seen in [24].

PCA is arguably the most widely used statistical tool for data analysis and dimensionality reduction today. However, its performance and applicability in real scenarios are limited by a lack of robustness to outlying or corrupted observations. Robust PCA (RPCA) recovers a low rank matrix L from corrupted observations $M = L+S$. The corrupted entries E are unknown and the errors can be arbitrarily large, but assumed to be sparse.

Robust principal component analysis (RPCA) proposed by Candes and colleagues [23]. Let $\|L\|_* = \Sigma_i\sigma_i(L)$ denote the nuclear norm of the matrix L, that is, the sum of its singular values, and let $\|S\|_1 = \Sigma_{ij}|S_{ij}|$ denote l_1-norm of S. Supposing that M denotes the observation matrix, RPCA solves the following optimization problem:

minimize $\|L\|_* + \lambda\|S\|_1$, subject to $M = L + S$

which exactly recovers the low-rank L_0 and the sparse S_0 with a positive regulation parameter λ. Theoretically, this is guaranteed to work even if the rank of L_0 grows almost linearly in the dimension of the matrix, and the errors in S_0 are up to a constant fraction of all entries. Algorithmically, we will see that the above problem can be solved by efficient and scalable algorithms, at a cost not so much higher than the classical PCA.

Lin *et al.* has given an algorithm solving the RPCA problem, which is referred to as the inexact ALM (IALM) method.

3.2.1 The RPCA Model of Gene Expression Data

We followed the same RPCA model for gene expression data, which is proposed in [12], except gene expression data weighted with literature information. Considering the matrix D of gene expression data weighted with literature information with size $m \times n$, each row of D represents the transcriptional responses of a gene weighted with literature information, in all the n samples, and each column of D represents the weighted expression levels of all the m genes in one sample. Without loss of generality, $m >> n$, so it is a classical small-sample-size problem.

Our goal of using RPCA to model the weighted microarray data is to identify these significant genes. Suppose the matrix decomposition $D = C + S$ has been done by using RPCA. By choosing the appropriate parameter λ , the sparse perturbation matrix S can be obtained, i.e. most of entries in S are zero. The genes corresponding to nonzero entries in S can be considered as ones of different expression.

The sparse matrix S can be denoted as:

$$S = \begin{bmatrix} s_{11} & s_{12} & .. & s_{1n} \\ s_{21} & s_{22} & .. & s_{2n} \\ ... & ... & ... & ... \\ s_{m1} & s_{m2} & .. & s_{mn} \end{bmatrix}$$

Then the following two steps are executed: firstly, the absolute values of entries in the sparse matrix S are find out; secondly, to get the evaluating vector, the matrix is summed by row. Consequently, the evaluating vector $\hat{S}$ is sorted in descending order. Without loss of generality, suppose that the first c_1 entries in $\hat{S}$ are non-zero, that is, $\hat{S} = [\hat{s_1}, \ldots, \hat{s_{c1}}, 0, \ldots, 0]^T$.

Generally, the larger the element in $\hat{S}$ is, the more different the gene is in expression data.

3.3 Least squares support vector machine classifiers

The constrained optimization problem for least squares support vector machine (LS-SVM) [25,26] for classification are defined as follows: Given a training set $\{x_i, y_i\}_{i=1}^N$ with input data $x_i \in \mathbb{R}^d$ and corresponding binary class labels $y_i \in \{-1, +1\}$. Vapnik's SVM classifier formulation was modified in [25] into the following LS-SVM formulation:

$$\min_{w,b,e} \frac{1}{2} w^T w + \gamma \frac{1}{2} \Sigma_{i=1}^N e_i^2$$

subject to:

$$y_i[w^T\varphi(x_i)+b] = 1 - e_i, \qquad i = 1, \ldots, N$$

where $\varphi(.)$: $\mathbb{R}^d \to \mathbb{R}^{d_h}$ is a nonlinear function which maps the d-dimensional input vector x from the input space to the d_h-dimensional feature space, possibly infinite. In the dual space the solution is given by

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + \frac{I}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix}$$

with $y = [y_1, \ldots, y_N]$, $1_N = [1, \ldots, 1]$, $e = [e_1, \ldots, e_N]$, $\beta = [\beta_1, \ldots, \beta_N]$ and $\Omega_{i,j} = y_i y_j K(x_i, x_j)$ where $K(x_i, x_j)$ is the kernel function.

The classifier in the dual space takes the form

$$y(x) = \text{sign}[\sum_{i=1}^{N} \beta_i y_i K(x, x_i) + b]$$

where β_i are Lagrange multipliers.

4 Results

In our analysis, we collected publically available binary class colon data set (disease vs. normal). Then we generated gene-term matrix as described in section data sets. In order to integrate these two data sets, we have applied the Equation (4). Thus colon cancer data set is weighted with literature information and resulted in a single matrix.

Then we applied RPCA on the integrated data set. In RPCA, the matrix has been decomposed as $D = A + S$. By choosing the appropriate parameter λ, the sparse perturbation matrix S can be obtained, i.e., most of entries in S are zero or near-zero. The genes corresponding to non-zero entries can be considered as differentially expressed ones.

For the sake of comparison, we performed prediction on the subsets of genes selected by RPCA in [12] and the proposed approach. Table 2 shows the classification performance of colon cancer data set on, whole data sets, subsets of genes identified by RPCA in [12] and proposed approach . The results show that better prediction performance in terms of test areas under curve (AUC) is obtained with the proposed approach. This emphasize that the identified differentially expressed genes are able to distinguish normal and cancerous samples quite well. Classification is performed with LS-SVM classifier in which 2/3rd samples are splitted into training set and remaining as the test sets. In all cases, linear kernel function is used with LS-SVM classifier. In addition, we assume that there are genes that are not selected due to the lack of literature information. So we searched for co-expressed genes of differentially expressed ones in [12] and proposed approach, with GeneMania [16]. We obtained subsets of co-expressed genes which are parts of colon cancer data, but not identified by our approach. Then we estimated

prediction performance on sets of genes, containing both co-expressed and differentially expressed ones. The predictive performances indicate that the incorporation of literature information into microarray analysis improve the identification of disease specific genes. In addition, searching for co-expressed genes help to identify disease related genes which are really missed in our analysis due to lack of literature information.

The colon cancer related genes were selected by the proposed approach are shown in Table 1. Several genes selected by this approach are known to be involved in, and important for, colon cancer. Ribosome, the essential cellular organelle for protein synthesis in all cells, consists of ribosomal RNAs (rRNAs) and ribosomal proteins (RPs). Ribosomal protein L41 (RPL41) is a microtubule-associated protein essential for functional spindles and for the integrity of centrosome and that the abnormal mitosis and disrupted centrosome associated with the RPL41 down-regulation may be related to malignant transformation [27]. In our analysis, RPL41 (H55933) ranked as the most differentially expressed gene. Studies in [28] and [29] already reported that RPL41 (H55933) gene has high rank in colon cancer.

Ectopic expression of tumor rejection antigen 1 (Tra1) were detected in the UC affected colonic mucosa [30]. Tra1 is reported as a differentially expressed gene in colon cancer [28]. Calcyclin binding protein (CacyBP) as a promising candidate biomarker for colorectal cancer (CRC) metastasis and also sheds light on the underlying molecular mechanism by which CacyBP promotes CRC metastasis [31]. 60S acidic ribosomal protein P1 (R02593) were reported as a top-ranked gene in colon cancer [28, 32].

Eukaryotic initiation factor (elF) 4A functions as a subunit of the initiation factor complex elF4F, which mediates the binding of mRNA to the ribosome. Serum response factor (SRF) regulates transcription of many serum-inducible and muscle-specific genes. It binds to the serum response element, a DNA sequence required for the transcription of a number of genes in response to growth factor or mitogen stimulation. A number of these types of genes are present in the top 50: b9-actin, myosin light chain, and profilin I. This is consistent with the observation that signal-regulated activation of SRF is mediated by changes in actin dynamics (46). These genes might provide an indication of the migratory capacity of the cells in the specimens and hence their propensity for metastasis [28, 33].

Human Leukocyte Antigen (HLA) class I molecules are of major importance for cell-mediated anti-tumor immune responses. Expression of HLA class I/2-microglobulin (2-m) complexes carrying tumor-specific peptides is a prerequisite for adaptively matured cytotoxic T cells (CTLs) to be able to recognize tumor cells [34]. Loss of expression of HLA class I molecules has been frequently reported for colorectal tumors [35, 36].

Low serum ferritin levels are associated with patients having serious gastrointestinal pathologies such as neoplasia and acid peptic disease [37]. Previous work has shown that the majority of colorectal adenocarcinomas exhibit

ferritin expression [38], but the clinical significance remains unknown. Myosin light chain alkali, smooth-muscle isoform (H20709) also reported as top ranked genes in colon cancer on several studies [28, 39].

Translationally controlled tumor protein (TCTP) is a highly conserved and ubiquitously expressed protein in all eukaryotes—highlighting its important functions in the cell. Previous studies revealed that TCTP is implicated in many biological processes, including cell growth, tumor reversion, and induction of pluripotent stem cell. In human colon cancer, the level of TCTP mRNA was detected in three human colon carcinoma cell lines (SNU-C2A, SNU-C4, and SNU-C5) [40]. Ornithine decarboxylase (OAZ1) catalyzes the conversion of ornithine to putrescine in the first and apparently rate-limiting step in polyamine biosynthesis. The ornithine decarboxylase antizymes play a role in the regulation of polyamine synthesis by binding to and inhibiting ornithine decarboxylase. OAZ(T63484) were reported [18, 28] as a top ranked gene in colon cancer.

Alterations in the distribution and/or adhesiveness of laminin receptors in colon cancer cell lines may be associated with increased tumorigenicity [41]. A study of cultured colon cancer cells suggests that laminin may play an important role in hematogeneous metastasis by mediating tethering and spreading of colon cancer cells under blood flow [42]. In general, the markers are involved in cell signaling, adhesion and communication, immune response, heat shock, and DNA repair [28]. Hepatocyte growth factor (HGF) signaling has been implicated in a broad spectrum of human cancers. Studies show that HGF-induced colon tumor cell proliferation, invasion as well as tumor growth and metastasis in xenograft models [43].

In short, out of 23 genes identified as differentially expressed of colon cancer, majority of these genes are reported as a top ranked genes on several studies. In addition, prediction performances on these selected genes clearly indicate that these gene are able to distinguish cancerous and non cancerous samples.

5 Discussion

The proposed method has been applied to colon cancer data of which the corresponding literature information is available. Literature information were gathered from Pubmed by searching gene and term co-occurrence. To verify the merit of our approach over the use of RPCA on single expression data source, models were built for classifying colon cancer patients. In many studies, single data sources are explored for the identification of differentially expressed genes. In our opinion, single data source is inadequate to explain complex network of genes underlying a disease. In this study, microarray data is weighted with literature information. These types of external knowledge into the analysis of microarray improves the identification of cancer related genes accurately.

In RPCA, firstly, the matrix S of differential expression was discovered from gene expression data matrix by using robust PCA; secondly, the differentially expressed genes were discovered according to matrix S . Thus it

provide an efficient and effective approach for gene selection. In addition, compared to other statistical approaches for gene selection, it does not widely vary on the number of genes selected on each iterations. But the parameter λ selection is important in order to obtain the optimal features.

For the data on colon cancer, the outcome - cancerous or not cancerous - could be predicted most accurately and most cost efficiently with an AUC ranged from 0.868 to 0.989. The biological relevance and the prediction performances of selected genes emphasize the relevance of our approach for the identification of differentially expressed genes.

6 Conclusion

The results suggest that the use of our integration approach on colon cancer data improve the identification of differentially expressed genes and performance of decision support in cancer. In real data examples, we have shown how to incorporate external knowledge, extracted from literature information on medical diagnosis. The proposed method provides a general way to incorporate such ever-increasing amounts of prior knowledge into the analysis and to further improve the predictive performance. These results emphasizes the need for comprehensive prior knowledge gathered with microarray data, but it is unknown that which type and levels of external knowledge are most relevant for the biomarker discovery and prognostic prediction.

7 Tables

Acknowledgements

BDM is full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, START 1, OT 09/052 Biomarker, several PhD/postdoc and fellow grants; Flemish Government: IOF: IOF/HB/10/039 Logic Insulin, IOF: HB/12/022 Endometriosis; FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits) research community MLDM; G.0733.09 (3UTR); G.0824.09 (EGFR); IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping; FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate); Hercules Stichting: Hercules III PacBio RS; iMinds 2013; IMEC: phd grant; EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED; COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network. The scientific responsibility is assumed by its authors.

Table 1: 23 Differentially expressed genes of colon cancer identified by proposed method

Gene Id		Gene Annotation
Hsa.3004	H55933	Homo sapiens mRNA for homologue to yeast ribosomal protein L41
Hsa.2357	T52342	Human tra1 mRNA for human homologue of murine tumor rejection antigen gp96
Hsa.474	L28809	Homo sapiens dbpB-like protein mRNA
Hsa.6080	J02763	Human calcyclin gene
Hsa.20836	R02593	60S acidic ribosomal protein P1
Hsa.45293	H86060	Negative factor (Simian immunodeficiency virus)
Hsa.20836	R02593	60S acidic ribosomal protein P1 (Polyorchis penicillatus)
Hsa.13491	R39465	Eukaryotic Initiation factor 4A (Oryctolagus cuniculus)
Hsa.3835	H79852	60S acidic ribosomal protein P2 (Babesia bovis)
Hsa.37254	R85482	Serum response factor (Homo sapiens)
Hsa.909	M11799	Human MHC class I HLA-Bw58 gene
Hsa.749	T63508	Ferritin in heavy chain (HUMAN)
Hsa.2597	T49423	Breast basic conserved protein 1 (HUMAN)
Hsa.750	T72863	Ferritin light chain (HUMAN)
Hsa.2800	X55715	Human Hums3 mRNA for 40S ribosomal protein s3
Hsa.8068	T57619	40S ribosomal protein S6 (Nicotiana tabacum)
Hsa.467	H20709	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN)
Hsa.3087	T65938	Translationally controlled tumor protein (HUMAN)
Hsa.5710	T63484	Human ornithine decarboxylase antizyme (Oaz) mRNA
Hsa.3061	X63469	Transcription initiation factor iie-beta chain (HUMAN)
Hsa.878	T61609	Laminin receptor (HUMAN)
Hsa.9994	T51539	Hepatocyte growth factor-like protein precursor (Homo sapiens)
Hsa.4689	T95018	40s ribosomal protein s18 (Homo sapiens)

Table 2: Comparison of classification accuracy on single data source and two data sources of colon cancer

Data Sources	No. of genes	Methods	Accuracy
Microarray and literature information	23	proposed approach	0.834(0.087)
Microarray	30	RPCA in [12]	0.834(0.085)
Microarray and literature information	33	proposed approach and GeneMania	**0.870**(0.078)
Microarray	46	RPCA in [12]and GeneMania	0.795(0.095)
Microarray	2000	whole data set	0.834(0.097)

References

1. Roth V, Lange T: **Bayesian Class Discovery in Microarray Data**. *IEEE Transactions on Biomedical Engineering* 2004, **51**.
2. Qiu P, Plevritis SK: **Simultaneous Class Discovery and Classification of Microarray Data Using Spectral Analysis**. *Journal of Computational Biology* 2009, **16**:935–944.
3. Somorjai RL, Dolenko B, Baumgartner R: **Class prediction and discovery using gene microarray and proteomics mass spectroscopy data:curses,caveats, cautions**. *Bioinformatics* 2003, **19**:1484–1491.
4. Conde L, Mateos A, Herrero J, J D: **Improved Class Prediction in DNA Microarray Gene Expression Data by Unsupervised Reduction of the Dimensionality followed by Supervised Learning with a Perceptron**. *Journal of VLSI Signal Processing* 2003, **35**:245–253.
5. Olga GT, Mitchell EG, Patrick OB, David B, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data**. *Bioinformatics* 2002, **18**:1454–1461.
6. Harriet W, Eeva K, Jouni KS, Antti K, Jaakko H, Sisko A, Sakari K: **Identification of differentially expressed genes in pulmonary adenocarcinoma by using cDNA array**. *Oncogene* 2002, **21**:5804–5813.
7. Sudhakar J, Rajagopalan S: **A PCA-Based approach for gene target selection to improve industrial strains**. *Computer Aided Chemical engineering* 2007, **24**:1013–1018.
8. Shao L, Huang Q, Luo M, Lai M: **Detection of the differentially expressed gene IGF-binding protein-related protein-1 and analysis of its relationship to fasting glucose in Chinese colorectal cancer patients**. *Endocrine-Related Cancer* 2004, **11**:141–148.
9. Chun LH, Wen CL: **Detecting differentially expressed genes in heterogeneous disease using half Student's t-test**. *Int.I.Epidemiol* 2010, **10**:1–8.
10. Sedeh S R, Bathe M, K-J B: **The Subspace Iteration Method in Protein Normal Mode Analysis**. *J Comput Chem* 2010, **31**:66–74.
11. Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genomescale expression data sets of two different organisms**. *PNAS* 2003, **100**:3351–3356.
12. Liu JX, Wang YT, Zheng CH, Sha W, Mi JX, Xu Y: **Robust PCA based method for discovering differentially expressed genes**. *BMC Bioinformatics* 2013, **14**.
13. Faro A, Giordano D, Spampinato C: **Combining literature text mining with microarray data: advances for system biology modeling**. *Brief Bioinform* 2012, **1**(13):61–82.
14. Cheung WA, Ouellette BF, Wasserman WW: **Inferring novel gene-disease associations using Medical Subject Heading Over-representation Profiles**. *Genome Medicine* 2012, (4):75.
15. Chaussabel D, Sher A: **Mining microarray expression data by literature profiling**. *Genome Biology* 2002, (3).
16. **GeneMania** [http://http://genemania.org/].
17. **Bioinformatics Research Group** [http://www.upo.es/eps/bigs/datasets.html].
18. Alon U, Barkai N, A Notterman D, Gish K, Ybarra S, Mack D, J Levine A: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays**. *PNAS* 1999, **96**:6745–6750.
19. **National Cancer Institute Dictionary of Cancer Terms** [http://www.cancer.gov/dictionary].
20. Golub GH, Van Loan CF: *Matrix Computations*. 2nd ed. (Baltimore: Johns Hopkins University Press) 1989.
21. Pearson K: **On Lines and Planes of Closest Fit to Systems of Points in Space**. *Philosophical Magazine 2* 1901, **11**:559–572.
22. Higham N: **Newton's Method for the Matrix Square Root**. *Mathematics of Computation* 1986, **46**(174):537–549.
23. Candes EJ, Li X, Ma Y, Wright J: **Robust Principal Component Analysis**. *Journal of the ACM* 2011, **58**(11).
24. Lin Z, Chen M, Wu L, Ma Y: **The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-rank Matrices**. *http://Arxiv.org/abs/1009.5055v2* 2010.
25. Suykens JAK, Vandewalle J: **Least Squares Support Vector Machine classifiers**. *Neural Processing Letters* 1999, **9**:293–300.

26. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J: *Least Squares Support Vector Machines*. Singapore: World Scientific 2002.

27. Wang S, Huang J, He J, Wang A, Xu S, Huang SF, Xiao S: **RPL41, a small ribosomal peptide deregulated in tumors, is essential for mitosis and centrosome integrity**. *Neoplasia* 2010, **3**(12):284–93.

28. J Moler E, L Chow M, S Mian I: **Analysis of molecular profile data using generative and discriminative methods**. *Physiol Genomics* 2000, **4**:109–126.

29. Li X, Rao S, Wang Y, Gong B: **Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling**. *Nucleic Acids Res* 2004, **9**(32):2685–2694.

30. Kurokawa MS, Hatsugai M, Noguchi Y, Yoshioka T, Mitsui H, Yasuda H, Kato T: *Proteomic Approaches for Biomarker Discovery in Ulcerative Colitis*. eBooks 2011.

31. Ghosh D, Li Z, Tan XF, Lim TK, Mao Y, Lin Q: **iTRAQ Based Quantitative Proteomics Approach Validated the Role of Calcyclin Binding Protein (CacyBP) in Promoting Colorectal Cancer Metastasis.** *Mol Cell Proteomics* 2013, **7**(12):1865–80.

32. Wang Z, Palade V: *Fuzzy gene mining: A fuzzy-based framework for cancer microarray data analysis*. A john wiley Sons 2009.

33. Sotiropoulos A, Gineitis D, Copeland J, Treisman R: **Signal-regulated activation of serum response factor is mediated by changes in actin dynamics**. *Cell* 1999, (98):159–169.

34. Townsend A, Bodmer H: **Antigen recognition by class I-restricted T lymphocytes**. *Annu Rev Immunol* 1989, (7):601–624.

35. Kaklamanis L, Gatter KC, Hil lAB, Mortensen N, Harris AL, Krausa P, McMichael A, Bodmer JG, Bodmer WF: **Loss of HLA class-I alleles, heavy chains and beta 2-microglobulin in colorectal cancer**. *Int J Cancer* 1992, (51):379–385.

36. Cabrera T, Collado A, Fernandez MA, Ferron A, Sancho J, Ruiz-Cabello F, Garrido F: **High frequency of altered HLA class I phenotypes in invasive colorectal carcinomas**. *Tissue Antigens* 1998, (52):114–123.

37. van den Elsen PJ, Holling TM, Kuipers HF, van der Stoep N: **Transcriptional regulation of antigen presentation**. *Curr Opin Immunol* 2004, (16):67–75.

38. Corver WE, Koopman LA, Mulder A, Cornelisse CJ, Fleuren GJ: **Distinction between HLA class I-positive and -negative cervical tumor subpopulations by multiparameter DNA flow cytometry**. *Cytometry* 2000, (41):73–80.

39. Huang TM, Kecman V, Kopriva I: *Kernel based algorithms for mining huge data sets: supervised, semi-supervised and unsupervised learning*. Springer 2006.

40. Chan THM, Chen L, Guan XY: **Role of translationally controlled tumor protein in cancer progression**. *Biochem Res Int* 2012, **2012**.

41. Kim WH, Lee BL, Jun SH, Song SY, Kleinman HK: **Expression of 32/67-kDa laminin receptor in laminin adhesionselected human colon cancer cell lines**. *Br J Cancer* 1998, (77):15–20.

42. Kitayama J, Nagawa H, Tsuno N, Osada T, Hatano K, Sunami E, Saito H, Muto T: **Laminin mediates tethering and spreading of colon cancer cells in physiological shear flow**. *Br J Cancer* 1999, (80):1927–1934.

43. Stein U, Walther W, Arlt F, Schwabe H, Smith J, Fichtner I, Birchmeier W, Schlag PM: **MACC1, a newly identified key regulator of HGF-MET signaling, predicts colon cancer metastasis**. *Nat Med* 2009, **1**(85):25–39.