International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:3, No:10, 2009

Oncogene Identification using Filter based Approaches between Various Cancer Types in Lung

Michael Netzer*, Michael Seger*, Mahesh Visvanathan*, Bernhard Pfeifer, Gerald H. Lushington, Christian Baumgartner

***Abstract*—**Lung cancer accounts for the most cancer related deaths for men as well as for women. The identification of cancer associated genes and the related pathways are essential to provide an important possibility in the prevention of many types of cancer. In this work two filter approaches, namely the information gain and the biomarker identifier (BMI) are used for the identification of different types of small-cell and non-small-cell lung cancer. A new method to determine the BMI thresholds is proposed to prioritize genes (i.e., primary, secondary and tertiary) using a k-means clustering approach. Sets of key genes were identified that can be found in several pathways. It turned out that the modified BMI is well suited for microarray data and therefore BMI is proposed as a powerful tool for the search for new and so far undiscovered genes related to cancer.

***Keywords*—**lung cancer, micro arrays, data mining, feature selection.

I. Introduction

LUNG cancer accounts for the most cancer related deaths (29%) for men as well as for women and follows with a very poor prognosis – a 5-year survival rate of 15% (data for USA) [1]. The major types of lung cancer are small-cell and non-small-cell cancer. Non-small-cell cancer can be further divided into three major histological subtypes: squamous-cell carcinoma, adenocarcinoma, and large-cell lung cancer [2]. The treatment of lung cancer depends on the cancer type and the stage of cancer including surgery, radiation therapy, chemotherapy and targeted biological therapies.

Biologists have known for a long time that the participation of certain genes in specific pathways are risk factors for multiple cancers. The identification of these genes and pathways is important since targeting them could provide an important possibility in the prevention of many types of cancer. Such genes include both onco-genes and onco-pathways that are amplified in cancers and activate the growth of tumors across different organs and tumor suppressor genes having the opposite effect (i.e., if active, they prevent multiple types of tumor growth and development) [3].

DNA microarray technology enables the simultaneous monitoring of the expression of thousands of genes resulting in a high dimensionality of the data subject to being investigated. Changes in the expression levels of single genes during cancer development within a given cell population may be associated with cancer etiology and development [4]. For extraction of those particular genes or features, however, sophisticated data mining approaches are required. Feature selection, as an important step in the data mining process, reduces dimensionality by searching for representative feature subsets with highly discriminatory ability.

In general, feature selection methods can be classified into filters and wrappers [5]. Filter methods rank features based on a quality measure (merit) depending on the ability to distinguish between predefined classes (e.g., case vs. control group). Wrappers use accuracy estimates provided by machine learning approaches to evaluate feature subsets. In general, feature subsets selected by wrappers are highly discriminatory, with the drawback of an extensive computational cost. Filters are more efficient but less accurate. The calculated merit – on the other hand – allows prioritizing features which is particularly important for biological interpretation purposes.

Especially for small size datasets there are significant differences in the ranking between different filter approaches due to the diversity of the underlying statistical models [6]. It is obvious that the underlying models learned from data include different types of errors. The bias-variance decomposition as defined by Geman and collegues [7] distinguishes between three types of errors: The *bias error* is a systematic component of the error. It results from differences between the learning method and the domain [8]. The *variance error* results from differences between models of different samples. The sum of bias and variance is called total expected error of a learning method. The *intrinsic error* is due to the uncertainty in the domain and cannot be “learned” [9].

In this work different types of small-cell and non-small-cell lung cancer are compared using information gain (IG) [10] and the biomarker identifier (BMI) [11] as feature selection methods. The IG computes the discriminatory ability of every feature based on an entropy measure.

The BMI, which was originally applied on metabolic data, combines various statistical measures to calculate an evaluation score for feature ranking. The strength of the BMI

*Authors contributed equally to this work
Michael Netzer, Michael Seger, Bernhard Pfeifer and Christian Baumgartner are with the Institute of Electrical and Bioengineering, University for Health Sciences, Medical Informatics and Technology (UMIT), A-6060 Hall in Tirol, Austria; e-mail: michael.netzer@umit.at.
Mahesh Visvanathan and Gerald H. Lushington are with the Bioinformatics Core Facility University of Kansas Lawrence, KS 66047, USA; e-mail: mvisvanathan@ku.edu

is the ability to clearly differ between primary, secondary and tertiary marker candidates with respect to their discriminatory ability. For the categorization of genes into these three groups using BMI a new method relying on a k-means clustering approach is proposed.

II. Methods

A. Research Data

In this study, the gene expression data sets from GlaxoSmithKline (GSK) are examined which had released the genomic profiling data for over 300 cancer cell lines via the National Cancer Institute's cancer Bioinformatics Grid™ (caBIG™) [12]. The investigated dataset applied in this work comprises data of 177 individuals divided into different types of lung cancer: small-cell (n = 41), adenocarcinoma (n = 65), squamous-cell (n = 34) and large-cell cancer (n = 37). Formally, the dataset can be described as a set of tuples *T*, where $T = \{(c_j,m) | c_j \in C, m \in M\}$ with C = { small-cell cancer, adenocarcinoma, squamous-cell, large-cell cancer }, C is the set of class labels and M is the set of features (gene expressions). The number of measured gene expressions available in the database is 54,676.

B. Feature Selection using the Information Gain

The IG describes how well a given feature separates between two or more classes based on an entropy measure. The IG with respect to class c_j can be defined as the difference between the entropy of class c_j and the conditional entropy for class c_j for a given feature f_i. This means that the expected reduction of entropy caused by partitioning the data according to feature f_i can be measured and used for feature ranking [10, 13]. More formally the IG in feature F with relation to C is the mutual information between F and C [14]:

$I(C,F) = H(C) - H(C \mid F)$, where

$H(C) = \sum_{c_i} P(c_i) \times \log_2 \frac{1}{P(c_j)}$, the initial entropy in C,

$H(C \mid F) = \sum_{f_i} P(f_i) \times H(C \mid f_j)$, the conditional entropy in C given F, and,

$H(C \mid f_j) = \sum_{c_i} P(c_i \mid f_j) \times \log_2 \frac{1}{P(c_j \mid f_j)}$, the entropy in C given a particular feature f_j.

C. Feature Selection using the Biomarker Identifier

The Biomarker identifier (BMI) was developed for dichotomous test problems and combines various statistical measures to discern the discriminatory ability of features distinguishing between two classes of interest. The BMI score for a feature f, a variant of the initial method described in Baumgartner et al. [11], is defined as:

$$BMI(f) = \lambda \cdot TP^2 \cdot \sqrt{\left|\Delta_{diff}\right| \frac{CV_{ref}}{CV}} \quad with$$

$$\Delta_{diff} = \begin{cases} \Delta & if \Delta \geq 1 \\ -\frac{1}{\Delta} & else \end{cases} \quad with \ \Delta = \frac{\bar{x}}{\bar{x}_{ref}}$$

where λ is a scaling factor and TP^2 is the product of the true positive (TP) values determined for both classes using logistic regression analysis. The parameter Δ_{diff} calculates relative changes in levels with respect to a reference group, and $\frac{CV_{ref}}{CV}$ denotes changes in the variance of data across the two cohorts. $\bar{x}$ is the mean value of levels in both classes. Using BMI for microarray data, a list of genes ranked by the BMI score is returned, representing the ability of genes to distinguish between both cohorts. Note that a positive Δ_{diff} can be interpreted as over expression, a negative Δ_{diff} value as under expression in the second class – compared to the chosen reference class – of a particular gene.

D. Gene categorization

A categorization scheme into primary, secondary and tertiary candidate genes according to their discriminatory ability is proposed. Primary genes reflect high (positive as well as negative) alterations in their expression levels. The prioritization into secondary and tertiary genes appears to be useful to distinguish between further promising candidates of which the latter group is more likely associated with secondary gene regulation pathways.

For the IG empirical threshold scores greater than zero, greater than the half maximum score and greater than two-thirds of the maximum score (see Table I) are used.

TABLE I. Thresholds for primary, secondary and tertiary gene sets using IG.

Categorization of genes	IG
Primary	≥ 0.67
Secondary	0.67 > IG ≥ 0.5
Tertiary	0.5 > IG > 0

To determine adequate thresholds for the BMI first a histogram of computed BMI scores was created (see Fig. 1). It is assumed that there are regions (or clusters) with "strong" (high absolute BMI score values, grey area in Fig. 1) and weak discriminating genes (low BMI absolute score values, black area in Fig. 1). To discern such regions a partitioning clustering algorithm on absolute BMI scores to get symmetric cut-offs was applied. In this work the k-means algorithm [15, 16] with k=4 number of clusters was used (three clusters represent genes categorized into primary, secondary and

tertiary genes, the cluster in the center of the histogram represents genes with weak or no discrimination). K-means groups the data objects by minimizing the sum of squared distances between each data point and its cluster representative based on an iterative procedure.

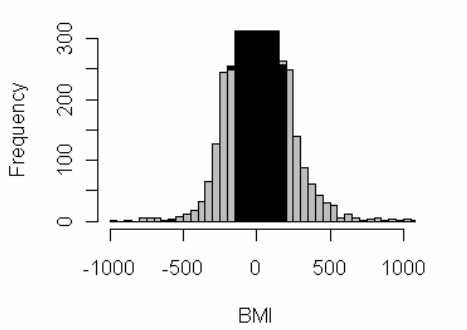

Fig. 1 Histogram of calculated BMI scores (schematic illustration). Grey areas indicate BMI scores of genes with good or excellent discrimination where the black area in the middle represents BMI values of genes with weak or no discrimination.

TABLE II. CALCULATED BMI THRESHOLDS FOR PRIMARY, SECONDARY AND TERTIARY GENE SETS.

Reference vs. comparison group	Categorization of genes		
	primary threshold	*secondary threshold*	*tertiary threshold*
Adenocarcinoma vs. small-cell	|BMI| ≥ 590	590 > |BMI| ≥ 170	170 > |BMI| ≥47
Squamous-cell vs. adenocarcinoma	|BMI| ≥ 230	230 > |BMI| ≥ 90	90 > |BMI| ≥ 30
Squamous-cell vs. large-cell	|BMI| ≥ 170	170 > |BMI| ≥ 50	50 > |BMI| ≥ 20
Squamous-cell vs. small-cell	|BMI| ≥ 190	190 > |BMI| ≥ 90	90 > |BMI| ≥ 40
Large-cell vs. adenocarcinoma	|BMI| ≥ 230	230 > |BMI| ≥ 80	80 > |BMI| ≥ 30
Large-cell vs. small-cell	|BMI| ≥ 140	140 > |BMI| ≥ 60	60 > |BMI| ≥ 30

III. RESULTS

The calculated BMI thresholds for gene categorization using the k-means approach are depicted in Table II. The corresponding clusters and thresholds for BMI when comparing adenocarcinoma vs. small-cell lung cancer are shown in Fig. 2. The identified number of primary, secondary and tertiary candidate genes using the BMI are depicted in Table III and using the IG in Table IV.

The IG lacked the ability to clearly categorize genes into the proposed scheme when using the clustering approach (Fig. 3), resulting in a high number of primary genes (2,531 primary gene candidates for adenocarcinoma vs. small-cell lung cancer, residual data not shown). This might be explained that IG scores do not follow roughly a Gaussian distribution (compare Fig. 3b). Furthermore the IG does not allow distinguishing between over- and under expression, because the IG solely delivers absolute values.

TABLE III. NUMBER OF IDENTIFIED GENES USING BMI AND K-MEANS CUT-OFFS FOR DIFFERENT LUNG CANCER TYPES.

Reference vs. comparison group	Categorization of genes		
	primary n	*secondary n*	*tertiary n*
Adenocarcinoma vs. small-cell	79	1669	13173
Squamous-cell vs. adenocarcinoma	321	4707	17464
Squamous-cell vs. large-cell	100	6121	34390
Squamous-cell vs. small-cell	614	6677	24028
Large-cell vs. adenocarcinoma	253	4981	16911
Large-cell vs. small-cell	555	11058	25771

TABLE IV. NUMBER OF IDENTIFIED GENES USING IG FOR DIFFERENT LUNG CANCER TYPES APPLYING THE THRESHOLDS DEFINED IN TABLE I.

Reference vs. comparison group	Categorization of genes		
	primary n	*secondary n*	*tertiary n*
Adenocarcinoma vs. small-cell	615	1503	21098
Squamous-cell vs. adenocarcinoma	13	519	26578
Squamous-cell vs. large-cell	0	20	5753
Squamous-cell vs. small-cell	78	993	25633
Large-cell vs. adenocarcinoma	13	519	26578
Large-cell vs. small-cell	78	993	25633

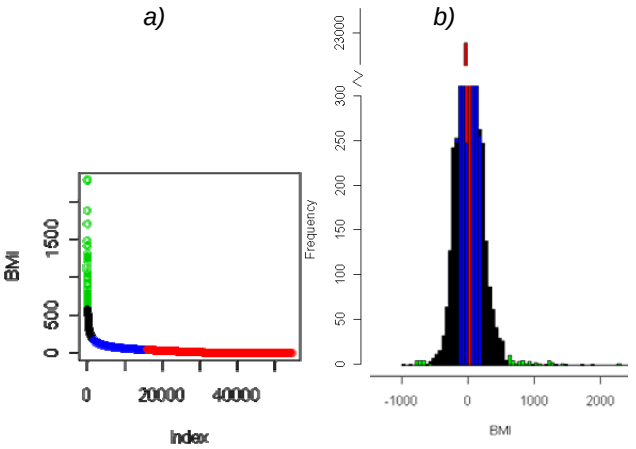

Fig. 2 Identified clusters on the BMI scores using the k-means algorithm for adenocarcinoma vs. small-cell lung cancer (a), and the related histogram plot (b). Green: primary genes; black: secondary genes; blue: tertiary genes. In the left figure (a) the absolute BMI scores are displayed according to their sorted rank (index).

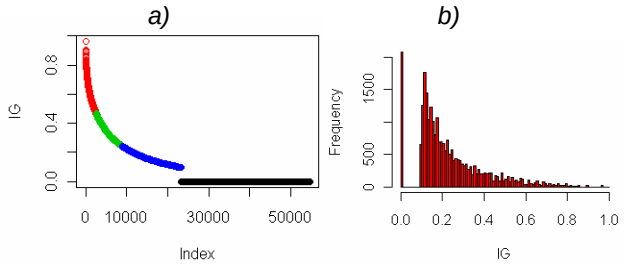

Fig. 3 Identified clusters on IG scores using the k-means clustering algorithm for adenocarcinoma vs. small-cell carcinoma (a) (red: primary genes; green: secondary genes; blue: tertiary genes), and the related histogram plot (b).

IV. Discussion and Conclusion

In this work gene expressions of different types of small-cell and non-small-cell lung cancer are compared. The feature selection methods IG and BMI were applied to search for the best discriminating genes when comparing pairs of different cancer types and categorize them into primary, secondary and tertiary candidate genes. It turned out that fixed thresholds are inappropriate for categorizing genes because the number of primary genes ranges from 0 to 615 for the different data sets when using empirical IG cut-offs. Based on this aspect a new method for adjusting thresholds using a k-means clustering approach was developed.

Due to the characteristics of roughly Gaussian distributed scores when using the BMI method it excellently turns out the primary gene cluster, representing a range beyond the 99th percentile of calculated BMI scores. Furthermore the BMI method is very useful to distinguish between over- and under expressed genes. Interpreting the distribution of BMI scores it also points out a general tendency to higher or lower over- or respectively under expressed genes in a micro array experiment (see Fig. 4).

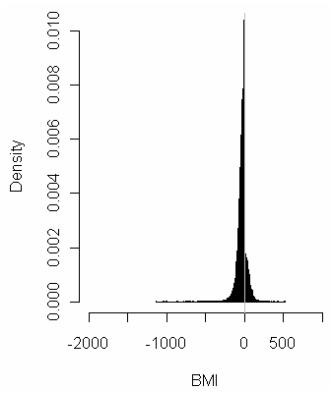

Fig. 4 Histogram plot of BMI scores for comparing squamous-cell vs. adenocarcinoma indicating a higher ratio of under expressed genes (BMI-scores < 0).

At this point it is also important to map the top ranking markers with their genes and biological pathways.

Therefore pathway analysis was performed using HGD (Hyper Geometric Distribution) technique to validate the top ranked genes and the associated pathways [17]. These genes were found to be part of *Cell Communication, Focal adhesion, T cell receptor signaling pathway, ECM-receptor interaction pathway, Cell Cycle and P53* signaling pathways. The list of top ranked gene names and their associated pathways by comparing squamous-cell lung carcinoma vs. large-cell lung carcinoma is shown in Table V (using BMI) and Table VI (using IG). Focal adhesion, cell cycle, P53 signaling and ECM-receptors pathways play a significant role in small cell lung cancer and non small cell lung cancer. These genes are involved in reducing the cell-cycle progression and degradation of resistance to apoptosis signals as observed in the small cell lung cancer pathway models. Genes like collagen, cyclin d1 that have been identified as one of the key genes are also responsible for constitutively up-regulation in lung cancer cell lines. They have been found to be ecteinascidin 743 (ET-743; Yondelis, Trabectedin) a marine anticancer agent that induced long-lasting objective remissions and tumor control in a subset of patients with lung carcinoma. Hence these primary genes identified through this approach can play a significant role in distinguishing various cancer types in lung.

TABLE V. Top ten ranked primary markers and pathways squamous-cell vs. large-cell using BMI.

Affymetrix ID	Gene Name	Pathways Involved
37892_at	collagen, type XI, alpha 1	Cell Communication, Focal adhesion, ECM-receptor interaction
242128_at	orthodenticle homolog 2	-
204320_at	collagen, type XI, alpha 1	Cell Communication, Focal adhesion, ECM-receptor interaction, Cell cycle
243610_at	otthump00000021439	-
206422_at	glucagon	-
1564359_a_at	similar to hypothetical protein FLJ36492	-
206378_at	n/a	-
219612_s_at	fibrinogen gamma chain	Complement Coagulation cascades, Small cell lung cancer
229271_x_at	n/a	-
210602_s_at	n/a	-

TABLE VI. TOP TEN RANKED PRIMARY MARKERS AND PATHWAYS SQUAMOUS-CELL VS. LARGE-CELL USING IG.

Affymetrix ID	Gene Name	Pathways Involved
217900_at	isoleucyl-tRNA synthetase 2, mitochondrial	Valine, leucine and isoleucine biosynthesis, Aminoacyl-tRNA biosynthesis
235072_s_at	n/a	-
211988_at	swi/snf related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	Chromatin Remodeling by hSWI/SNF ATP-dependent Complexes, Control of Gene Expression by Vitamin D Receptor
218820_at	chromosome 14 open reading frame 132	-
209177_at	chromosome 3 open reading frame 60	-
208711_s_at	cyclin d1	Cell cycle, p53 signaling pathway, Wnt signaling pathway, Focal adhesion, Small cell lung cancer, Non-small cell lung cancer
212614_at	at rich interactive domain 5b (mrf1-like)	-
226609_at	discoidin, cub and lccl domain containing 1	-
222572_at	protein phosphatase 2c, magnesium-dependent, catalytic subunit	-
218754_at	nucleolar protein 9	-

In order to cross validate the findings based on the selected cell lines used in the caBIG™ gene expression studies should be probed for expression profile of the identified genes and the corresponding protein levels. Similar profiling of the tumor tissues from mouse and human tumors would further validate findings from BMI and IG. An additional level of validation could involve pharmacologically treating the cells with known anti-tumor agents and profiling the same genes to determine potential efficacy. The biological studies might confirm the accuracy of the informatics tools developed and also point toward selective biomarkers that may be of significance in diagnostic and prognostic applications.

Using IG and BMI sets of key genes which can be found in several pathways could be identified. Especially the BMI combined with dynamic thresholds is well suited for analyzing microarray experiments and therefore BMI as a powerful tool for the exploration of new and so far undiscovered genes associated with cancer is proposed.

For future work it is intended to further study the predictive value of discovered gene sets to aid in risk prediction in lung cancer.

ACKNOWLEDGMENT

This publication was made possible by grant number P20 RR016475 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and the Austrian GEN-AU project Bioinformatics Integration Network.

REFERENCES

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray and M.J. Thun, "Cancer Statistics", *CA Cancer J Clin*, vol 58, pp. 71-96, 2008.

[2] R.S. Herbst, J.V. Heymach, S.M. Lippman, "Lung cancer." , *N Engl J Med*., vol. 360, pp. 87-8, 2009.

[3] I.G. Campbell, S.E. Russell, D.Y. Choong, K.G. Montgomery, M.L. Ciavarella, C.S. Hooi, B.E. Cristiano, P.B. Pearson, W.A. Phillips, "Mutation of the pik3ca gene in ovarian and breast cancer", *Cancer Res*., vol. 64, pp. 7678–7681, 2004.

[4] R. Hewett and P. Kijsanayothin, "Tumor classification ranking from microarray data", *BMC Genomics*, vol. 9, 2008.

[5] C. Baumgartner and A. Graber, "Data mining and knowledge discovery in metabolomics," In Masseglia F, Poncelet P, Teisseire M (eds.) *Successes and new directions in data mining*. Idea Group Inc, 2007, pp. 141-166.

[6] M. Netzer, G. Millonig, M. Osl, B. Pfeifer, S. Praun, J. Villinger, W. Vogel and C. Baumgartner, "A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry (IMR-MS)", *Bioinformatics*, vol. 25, pp. 941-947, 2009.

[7] S. Geman, E. Bienenstock and R. Doursat, "Neural networks and the bias/variance dilemma.", *Neural Computation*, vol. 4, pp. 1-58, 1992.

[8] P. Putten and M. Someren, "A bias-variance analysis of a real world learning problem: the coil challenge 2000." *Machine Learning*, vol. 57, pp. 177-195, 2004.

[9] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[10] R.J. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann, 1993.

[11] C. Baumgartner and D. Baumgartner, "Biomarker discovery, disease classification, and similarity query processing on high-throughput ms/ms data of inborn errors of metabolism." *J Biomol Screen*, vol. 11, pp. 90-99, 2006.

[12] NCI, *https://array.nci.nih.gov/caarray/project/details.action?project.experiment.publicIdentifier=woost-00041#*; last visited on April 9th, 2009.

[13] M. Osl, S. Dreiseitl, B. Pfeifer, K. Weinberger, H. Klocker, G. Bartsch, G. Schäfer, B. Tilg, A. Graber, and C. Baumgartner, "A new rule-based data mining algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry." *Bioinformatics*, vol. 24, pp. 2908-2914, 2008.

[14] J.D. Nelson, "Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain." *Psychol Rev*., pp. 979-99, 2005.

[15] J.B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

[16] J.A. Hartigan and M.A. Wong, "A k-means clustering algorithm." *JR Stat. Soc. Ser. C-Appl. Stat*, *28*:100-108, 1979.

[17] R. Barriot, J. Poix., A. Groppi, A. Barre., N. Goffard., D. Sherman., I. Dutour and A. de Daruvar, "New strategy for the representation and the integration of biomolecular knowledge at a cellular scale." *Nucleic Acids Res*., vol. 32, pp. 3581–3589, 2004.