Journal of Experimental & Clinical Cancer Research

BioMed Central

Research

Open Access

Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method

Peng Guan[1,2], Desheng Huang[1,2], Miao He[3] and Baosen Zhou*[1,2]

Address: [1]Department of Epidemiology, School of Public Health, China Medical University, Shenyang 110001, PR China, [2]Key Laboratory of Cancer Etiology and Intervention, University of Liaoning Province, Shenyang 110001, PR China and [3]Information Center, the First Affiliated Hospital, China Medical University, Shenyang 110001, PR China

Email: Peng Guan - pguan@mail.cmu.edu.cn; Desheng Huang - dshuang@mail.cmu.edu.cn; Miao He - job-mail@263.net; Baosen Zhou* - bszhou@mail.cmu.edu.cn

* Corresponding author

Published: 18 July 2009

Journal of Experimental & Clinical Cancer Research 2009, **28**:103 doi:10.1186/1756-9966-28-103

Received: 3 June 2009
Accepted: 18 July 2009

This article is available from: http://www.jeccr.com/content/28/1/103

Abstract

Background: A reliable and precise classification is essential for successful diagnosis and treatment of cancer. Gene expression microarrays have provided the high-throughput platform to discover genomic biomarkers for cancer diagnosis and prognosis. Rational use of the available bioinformation can not only effectively remove or suppress noise in gene chips, but also avoid one-sided results of separate experiment. However, only some studies have been aware of the importance of prior information in cancer classification.

Methods: Together with the application of support vector machine as the discriminant approach, we proposed one modified method that incorporated prior knowledge into cancer classification based on gene expression data to improve accuracy. A public well-known dataset, Malignant pleural mesothelioma and lung adenocarcinoma gene expression database, was used in this study. Prior knowledge is viewed here as a means of directing the classifier using known lung adenocarcinoma related genes. The procedures were performed by software R 2.80.

Results: The modified method performed better after incorporating prior knowledge. Accuracy of the modified method improved from 98.86% to 100% in training set and from 98.51% to 99.06% in test set. The standard deviations of the modified method decreased from 0.26% to 0 in training set and from 3.04% to 2.10% in test set.

Conclusion: The method that incorporates prior knowledge into discriminant analysis could effectively improve the capacity and reduce the impact of noise. This idea may have good future not only in practice but also in methodology.

Background

A reliable and precise classification is essential for successful diagnosis and treatment of cancer. Thus, improvements in cancer classification have attracted more attention [1,2]. Current cancer classification is mainly based on clinicopathological features, gene expression microarrays have provided the high-throughput platform to discover genomic biomarkers for cancer diagnosis and

prognosis [3-5]. Microarray experiments also led to a more complete understanding of the molecular variations among tumors and hence to a more accurate and informative classification [6-9]. However, this kind of knowledge is often difficult to grasp, and turning raw microarray data into biological understanding is by no means a simple task. Even a simple, small-scale, microarray experiment generates thousands to millions of data points.

Current methods to help classifying human malignancies based on microarray data mostly rely on a variety of feature selection methods and classifiers for selecting informative genes [10-12]. The ordinary process of gene expression data is as follows: first, a subset of genes with known classification is randomly selected (training set), then, the classifier is trained in the above training set until it is mature, finally, the classifier is used to perform the classification of unknown gene expression data. Commonly employed methods of feature gene selection included Nearest Shrunken Centroids (also known as prediction analysis for microarrays, PAM), shrunken centroids regularized discriminant analysis (SCRDA) and multiple testing procedure(MTP). The conventional methods of classification included k nearest-neighbor classifiers(KNN), linear discriminant analysis(LDA), support vector machine(SVM), back-propagation artificial neural network(BP-ANN) and etc, while the choice of which is a matter of dispute among methodologists [13-15]. So, improvement of existing methods or development of new methods is needed for the analysis of gene expression microarray data. Many gene expression signatures have been identified in recent years for accurate classification of tumor subtypes [16-19]. It has been indicated that rational use of the available bioinformation can not only effectively remove or suppress noise in gene chips, but also avoid one-sided results of separate experiment. However, a relatively few attempts have been aware of the importance of prior information in cancer classification [20-22].

Lung cancer is one of the leading causes of cancer death worldwide [23-26], can be classified broadly into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), and adenocarcinoma is the most common form of lung cancer. Because in China the cigarette smoking rate continues to be at a high level [27], a peak in lung cancer incidence is still expected [28]. Therefore, only lung cancer gene expression microarray dataset was selected in the present study.

In summary, together with the application of support vector machine as the discriminant approach and PAM as the feature gene selection method, we propose one method that incorporates prior knowledge into cancer classification based on gene expression data. Our goal is to improve classification accuracy based on the publicly available lung cancer microarray dataset [29].

Methods

Microarray dataset

In the present study, we analyzed the well-known and publicly available microarray dataset, malignant pleural mesothelioma and lung adenocarcinoma gene expression database http://www.chestsurg.org/publications/2002-microarray.aspx[29]. This Affymetrix Human GeneAtlas U95Av2 microarray dataset contains 12 533 genes' expression profiles of 31 malignant pleural mesothelioma (MPM) and 150 lung adenocarcinomas (ADCA, published in a previous study [30]), aims to test expression ratio-based analysis to differentiating between MPM and lung cancer. In this dataset, a training set consisted of 16 ADCA and 16 MPM samples.

Microarray data preprocessing

The absolute values of the raw data were used, then they were normalized by natural logarithm transformation. This preprocessing procedure was performed by using R statistical software version 2.80 (R foundation for Statistical Computer, Vienna, Austria).

Gene selection via PAM

Prediction analysis for microarrays (PAM, also known as Nearest Shrunken Centroids) is a clustering technique used for classification, it uses gene expression data to calculate the shrunken centroid for each class and then predicts which class an unknown sample would fall into based on the nearest shrunken centroid. Through this process, it can also identify the specific genes that most determine the centroid. The details of PAM method can be found in several published studies [31,32]. Here we adopted ten independent repeats of 10-fold cross-validation (CV) to avoid overlapping test sets. First, the preprocessed dataset was split into 10 subsets of approximately equal size by random sampling, secondly, each subset in turn was used for testing and the remaining 9 subsets for training. The above procedure was repeated 10 times. The error estimates were averaged to yield an overall error estimate. Note that the training set included 100 samples (16290 cases) and the test set included 100 samples (1810 cases) after the above ten independent repeats of 10-fold cross-validation.

Gene selection via prior biological knowledge

Published studies were collected in the database National Library of Medicine on the web (http://www.ncbi.nlm.nih.gov/sites/entrez, Pubmed) from Jan 1st, 2000 until March 31st, 2009 according to the retrieval strategy of "human lung adenocaicinoma" and published in the journal entitled "Cancer Research". Prior knowledge was viewed here as a means of directing the classifier

using known lung adenocarcinoma genes. For the purposes of this study, prior knowledge was any information about lung adenocarcinoma related genes that have been confirmed in literature. Hence, due to the journal's scope and the author's institution's accessibility, we restricted our attention to the journal entitled "Cancer Research". Cancer Research's publication scope covers all subfields of cancer research. The full texts of the papers were downloaded and then lung adenocarcinoma-related genes were retrieved from the literature. Then, after these genes' locations in the original dataset were collected, the genes were tested through multiple testing procedure in the training set provided by Gordon et al [29]. Significant genes were retained after the significant level was set as 0.05 to exclude the non-significant genes.

The combination of the feature genes selected by PAM method and from prior knowledge will be used to direct following classification.

Classification via modified SVM

Support Vector Machines (SVM) developed by Cortes & Vapnik [33] in 1995 for binary classification is currently a hot topic in the machine learning theory and one of the most powerful techniques for classification of microarray data. SVM's basic idea for classification may be roughly shown as follows, basically, we are looking for the optimal separating hyperplane between the two classes by maximizing the margin between the classes' closest points (see Figure 1) – the points lying on the boundaries are called support vectors H_1 and H_2, and the middle of the margin H is the optimal separating hyperplane. Except for linear decision making, SVM can also solve non-linear problems by first mapping the data to some higher dimensional feature space and constructing a separating hyperplane in this space. Several kernel functions have been introduced in order to deal with non-linear decision surfaces, (1) linear kernel: $K(x, y) = x \bullet y$; (2) polynomial kernel: $K(x, y) = [(x \bullet y)+c]^d$, d = 1, 2, ...; (3) radial basis kernel: $K(x, y) = \exp\{-|x-y|^2/\sigma^2\}$; (4) Sigmoid kernel: $K(x, y) = \tanh[b(x \bullet y)+c]$, where b, c and σ are parameters. Among these four types of kernel function, radial basis kernel showed best performance according to the results from similar studies [34,35]. The correct choice of kernel parameters is crucial for obtaining good results, so an extensive search must be conducted on the parameter space before results can be trusted. Here we adopted radial basis kernel function and 5-fold cross-validation in the training set to search the best parameters for SVM-based classification in the test set.

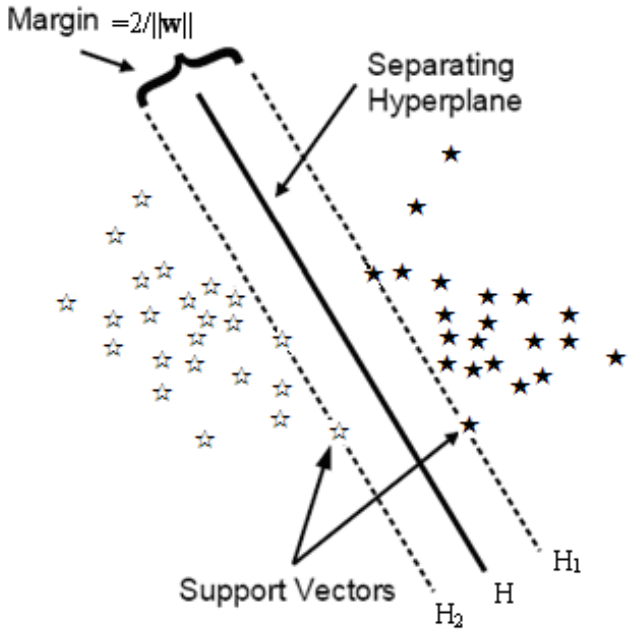

Figure 1
Classification via SVM (linear separable case).

Evaluation of model performance

Classification accuracy and the standard deviations of our proposed method (with prior knowledge) were compared with the original one (no prior knowledge) in the training set and test set. The framework of the above mentioned procedures is shown in Figure 2.

Statistical analysis

All the statistical analyses were conducted using R statistical software version 2.80 (R foundation for Statistical Computer, Vienna, Austria).

Results

Genes selected by PAM

The number of genes selected by PAM method varied from 4 to 12 with an average 7.81, and the standard deviation 2.21. The combination of genes selected by PAM is shown in Table 1. Among them, CEACAM6, calretinin, VAC-β and TACSTD1 appeared in the results all the time.

Gene selection via prior biological knowledge

After reviewed the full text of literature, twenty-three lung adenocarcinoma-related genes were selected. Then, Table 2 lists the eight significant genes that passed the multiple testing procedure in the training set provided by Gordon et al. The details of these genes are shown in Table 2.

Evaluation of model performance

Our proposed method performed better after incorporating prior knowledge (Figure 3). Accuracy of the modified method improved from 98.86% to 100% in training set and from 98.51% to 99.06% in test set. The standard devi-

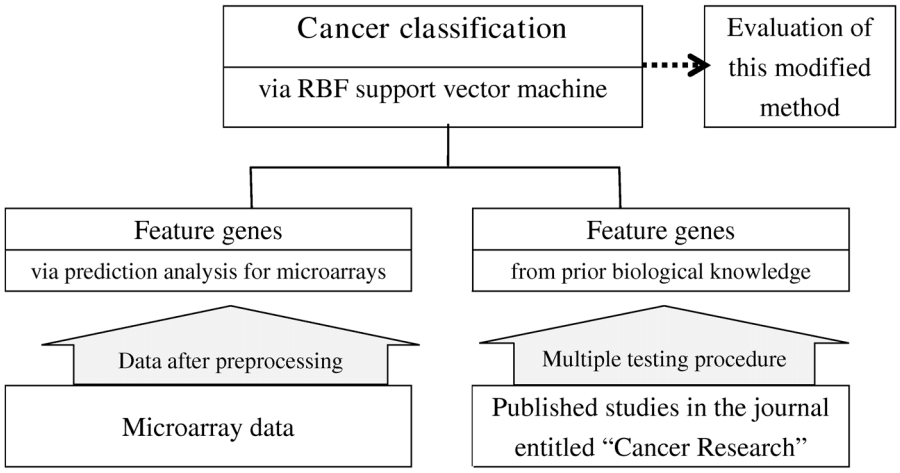

Figure 2
Framework of our proposed method.

ation of the modified method decreased from 0.26% to 0 in training set and from 3.04% to 2.10% in test set.

Here, we considered another situation, if there was an overlap between the two sources of genes, i.e. there existed the multi-collinearity, was there any influence on the performance of classification? Hence, taking into account the effect of overlap seemed natural for the current study. Expression quantity of VAC-β with a coefficient 1, 0.5 and 0.05 which meant complete, strong and minor correlation was added to data set for comparison, respectively. The accuracy in the above situation is 99.12%, 99.28%, 99.23% with the standard deviation 2.04%, 2.04%, 1.93%, respectively (Figure 3). McNemar's test was adopted to compare the accuracy between 'no prior knowledge' and the other 4 situations (with prior knowledge, complete correlation with prior knowledge, strong correlation with prior knowledge and minor correlation with prior knowledge) in training set and test set, and all the differences were statistically significant.

The accuracy in the training set was better than that in the test set, and the standard deviations were lower in training set than those in test set. Although Chi-square test indicated that the differences between them were statistically significant, the two sets were not comparable, and the difference may be caused by the large sample size. Training set was used for training and fitting, while test set focused on testing the ability to extrapolate.

Table 1: Gene lists selected by Prediction Analysis for Microarrays

Gene name	GenBank access No.	Location at HG_U95Av2
ERBB3	M34309	1585_at
CD24	L33930	266_s_at
TACSTD2	J04152	291_s_at
UPK1B	AB015234	32382_at
HIST1H2BD	M60751	38576_at
TITF-1	U43203	33754_at
CLDN3	AB000714	33904_at
CEACAM6	M18728	36105_at
PTGIS	D83402	36533_at
SFTPB	J02761	37004_at
caltrtinin	X56667	37157_at
VAC-β	X16662	37954_at
claudin-7	AJ011497	38482_at
AGR2	AF038451	38827_at
TACSTD1	M93036	575_s_at

Discussion

Microarrays are capable of determining the expression levels of thousands of genes simultaneously and have greatly facilitated the discovery of new biological knowledge [36]. One feature of microarray data is that the number of tumor samples collected tends to be much smaller than the number of genes. The number for the former tends to

Table 2: Genes as prior biological knowledge

Gene name	GenBank access No.	Location at HG_U95Av2
CXCL1	J03561	408_at
IL-18	U90434	1165_at
AKAP12	X97335	37680_at
KLF6	U51869	37026_at
AXL	M76125	38433_at
MMP-12	L23808	1482_g_at
PKP3	Z98265	41359_at
CYP2A13	U22028	1553_r_at

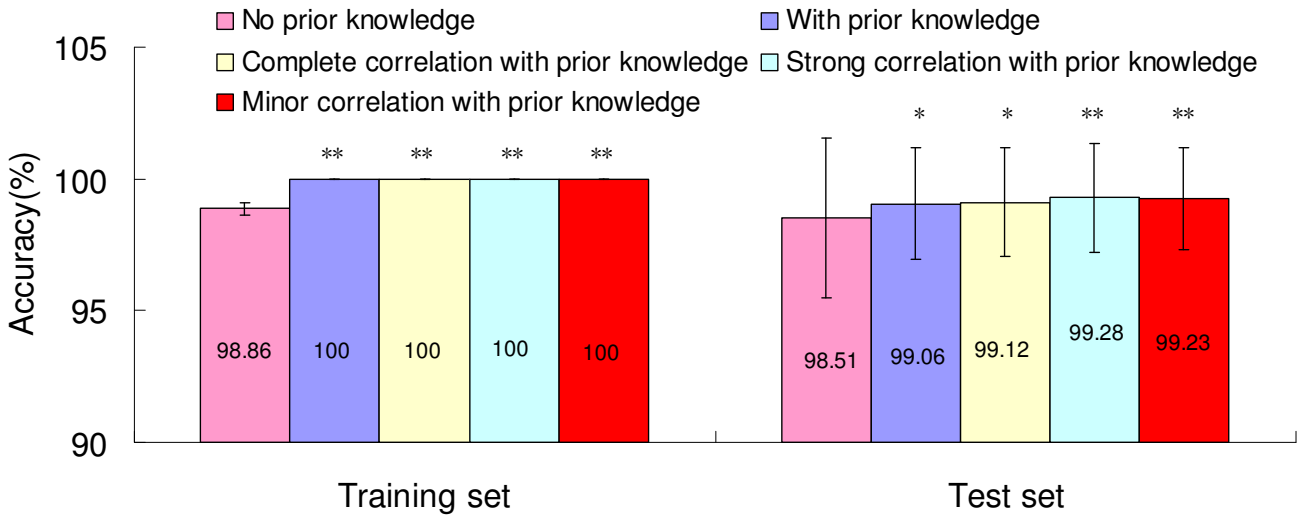

Figure 3
Accuracy comparisons, no prior knowledge vs. with prior knowledge. Note: * Accuracy is significantly higher when compared to no prior knowledge at the 0.05 level (2-tailed). ** Accuracy is significantly higher when compared to no prior knowledge at the 0.01 level (2-tailed).

be on the order of tens or hundreds, while microarray data typically contain thousands of genes on each chip. In statistical terms, it is called 'large p, small n' problem, i.e. the number of predictor variables is much larger than the number of samples. Thus, microarrays present new challenge for statistical methods and improvement of existing statistical methods is needed. Our research group's interest is lung cancer, we found that one of the key issues in lung cancer diagnosis was the discrimination of a primary lung adenocarcinoma from a distant metastasis to the lung, and so, it was important to identify which contribute most to the classification.

The present study used the combination of the genes selected by PAM and the genes from published studies, the result of this proposed idea was superior to that only rely on the genes selected by PAM. Considered from the methodological point, if the priori knowledge is not contrary to the truth, the incorporation of priori information is able to improve the classification accuracy, at least can not reduce the performance. From the point of accuracy improvement, our result is of concordance with the results of other previous studies [37,38]. It is interesting to compare the list of 15 genes selected by PAM and 8 genes as prior biological knowledge. In the current study, there was no overlap between these two gene lists, but the situation of overlap may be encountered in practice. Several genes may share the same or similar functions, so the existing of correlations among these genes from these two sources should be considered. Our result indicated that after the correlated gene had been added, no decrease of accuracy was found, which meant that there was no need to pay excess attention to the situation that overlapping existed between the information from microarray data and prior information.

One of the main limitations for the present study was how to incorporate prior biological knowledge and where to get it from. The prior biological knowledge in our study was retrieved from the literature, while, with the development of science and technology, huge knowledge will be discovered and reported. The magnitude of prior knowledge may have a certain impact on the results more or less. What information can be used as the truth and which kind of information should be excluded need to be further explored, maybe some experience could be borrowed from evidence-based medicine. On the other hand, the minimum number of predictor genes is not known, which may serve as a potential limitation of the study, and the discrimination function can vary (for the same genes) based on the location and protocol used for sample preparation [39]. The complexity of discriminant analysis and the multiple choices among the available discriminant methods are quite difficult tasks, which may influence the adoption by the clinicians in the future. Although highly accurate, microarray data's widespread clinical relevance and applicability are still unresolved.

Conclusion

In summary, a simple and general framework to incorporate prior knowledge into discriminant analysis was proposed. Our method seems to be useful for the

improvement of classification accuracy. This idea may have good future not only in practice but also in methodology.

Abbreviations

PAM: prediction analysis for microarrays; SCRDA: shrunken centroids regularized discriminant analysis; MTP: multiple testing procedure; KNN: k nearest-neighbor classifiers; LDA: linear discriminant analysis; SVM: support vector machine; BP-ANN: back-propagation artificial neural network; SCLC: small cell lung cancer; NSCLC: non-small cell lung cancer; MPM: malignant pleural mesothelioma; ADCA: adenocarcinoma; CV: cross-validation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PG conceived the study and drafted the manuscript. PG, DH, MH and BZ retrieved and reviewed the literature. PG and BZ attracted funding. All authors contributed to the writing of the final version of this paper.

Acknowledgements

This study was partially supported by Provincial Education Department of Liaoning (No.2008S232), Natural Science Foundation of Liaoning province (No.20072103) and China Medical Board (No.00726.). The authors are most grateful to the contributors of the dataset and R statistical software. Peng Guan was supported by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (No. [2008]890) and a CMU Development grant (No. [2008]5).

References

1. Lancashire LJ, Lemetre C, Ball GR: **An introduction to artificial neural networks in bioinformatics – application to complex microarray and mass spectrometry datasets in cancer studies.** *Brief Bioinform* 2009, **10:**315-329.
2. Liao JG, Chin KV: **Logistic regression for disease classification using microarray data: model selection in a large p and small n case.** *Bioinformatics* 2007, **23:**1945-1951.
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
4. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S: **Gene-expression profiles predict survival of patients with lung adenocarcinoma.** *Nat Med* 2002, **8:**816-824.
5. Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nat Genet* 2003, **33:**49-54.
6. Chen PC, Huang SY, Chen WJ, Hsiao CK: **A new regularized least squares support vector regression for gene selection.** *BMC Bioinformatics* 2009, **10:**44.
7. Statnikov A, Wang L, Aliferis CF: **A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.** *BMC Bioinformatics* 2008, **9:**319.
8. Boulesteix AL, Porzelius C, Daumer M: **Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value.** *Bioinformatics* 2008, **24:**1698-1706.
9. Baker SG, Kramer BS: **Identifying genes that contribute most to good classification in microarrays.** *BMC Bioinformatics* 2006, **7:**407.
10. Liu Z, Tan M, Jiang F: **Regularized F-measure maximization for feature selection and classification.** *J Biomed Biotechnol* 2009, **2009:**617946.
11. Lee YJ, Chang CC, Chao CH: **Incremental forward feature selection with application to microarray gene expression data.** *J Biopharm Stat* 2008, **18:**827-840.
12. Chen Z, Li J, Wei L: **A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue.** *Artif Intell Med* 2007, **41:**161-175.
13. Yousef M, Jung S, Showe LC, Showe MK: **Recursive cluster elimination (RCE) for classification and feature selection from gene expression data.** *BMC Bioinformatics* 2007, **8:**144.
14. Wu W, Xing EP, Myers C, Mian IS, Bissell MJ: **Evaluation of normalization methods for cDNA microarray data by k-NN classification.** *BMC Bioinformatics* 2005, **6:**191.
15. Laderas T, McWeeney S: **Consensus framework for exploring microarray data using multiple clustering methods.** *OMICS* 2007, **11:**116-128.
16. Botting SK, Trzeciakowski JP, Benoit MF, Salama SA, Diaz-Arrastia CR: **Sample entropy analysis of cervical neoplasia gene-expression signatures.** *BMC Bioinformatics* 2009, **10:**66.
17. Abba MC, Sun H, Hawkins KA, Drake JA, Hu Y, Nunez MI, Gaddis S, Shi T, Horvath S, Sahin A, Aldaz CM: **Breast cancer molecular signatures as determined by SAGE: correlation with lymph node status.** *Mol Cancer Res* 2007, **5:**881-890.
18. Xu L, Geman D, Winslow RL: **Large-scale integration of cancer microarray data identifies a robust common cancer signature.** *BMC Bioinformatics* 2007, **8:**275.
19. Fu LM, Fu-Liu CS: **Multi-class cancer subtype classification based on gene expression signatures with reliability analysis.** *FEBS Lett* 2004, **561:**186-190.
20. Chen X, Wang L: **Integrating biological knowledge with gene expression profiles for survival prediction of cancer.** *J Comput Biol* 2009, **16:**265-278.
21. Tai F, Pan W: **Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data.** *Bioinformatics* 2007, **23:**3170-3177.
22. Le Phillip P, Bahl A, Ungar LH: **Using prior knowledge to improve genetic network reconstruction from microarray data.** In *Silico Biol* 2004, **4:**335-353.
23. Karim-Kos HE, de Vries E, Soerjomataram I, Lemmens V, Siesling S, Coebergh JW: **Recent trends of cancer in Europe: A combined approach of incidence, survival and mortality for 17 cancer sites since the 1990s.** *Eur J Cancer* 2008, **44:**1345-1389.
24. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA: **Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship.** *Mayo Clin Proc* 2008, **83:**584-594.
25. Tyczynski JE, Bray F, Aareleid T, Dalmas M, Kurtinaitis J, Plesko I, Pompe-Kirn V, Stengrevics A, Parkin DM: **Lung cancer mortality patterns in selected Central, Eastern and Southern European countries.** *Int J Cancer* 2004, **109:**598-610.
26. Janssen-Heijnen ML, Coebergh JW: **The changing epidemiology of lung cancer in Europe.** *Lung Cancer* 2003, **41:**245-58.
27. Gu D, Kelly TN, Wu X, Chen J, Samet JM, Huang JF, Zhu M, Chen JC, Chen CS, Duan X, Klag MJ, He J: **Mortality attributable to smoking in China.** *N Engl J Med* 2009, **360:**150-159.
28. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA: **Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship.** *Mayo Clin Proc* 2008, **83:**584-594.
29. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R: **Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma.** *Cancer Res* 2002, **62:**4963-4967.
30. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci USA* 2001, **98:**13790-13795.

31. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays.** *Stat Sci* 2003, **18:**104-117.
32. Wang S, Zhu J: **Improved centroids estimation for the nearest shrunken centroid classifier.** *Bioinformatics* 2007, **23:**972-979.
33. Cortes C, Vapnik V: **Support-vector network.** *Mach Learn* 1995, **20:**1-25.
34. Pirooznia M, Yang JY, Yang MQ, Deng Y: **A comparative study of different machine learning methods on microarray gene expression data.** *BMC Genomics* 2008, **9(Suppl 1):**S13.
35. Pirooznia M, Deng Y: **SVM Classifier-a comprehensive java interface for support vector machine classification of microarray data.** *BMC Bioinformatics* 2006, **7(Suppl 4):**S25.
36. Campioni M, Ambrogi V, Pompeo E, Citro G, Castelli M, Spugnini EP, Gatti A, Cardelli P, Lorenzon L, Baldi A, Mineo TC: **Identification of genes down-regulated during lung cancer progression: a cDNA array study.** *J Exp Clin Cancer Res* 2008, **27:**38.
37. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information.** *Bioinformatics* 2005, **21:**2988-2993.
38. Huang D, Pan W: **Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data.** *Bioinformatics* 2006, **22:**1259-1268.
39. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Landers ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.