Published in final edited form as:
Curr Bioinform. 2014 April 1; 9(2): 140–146. doi:10.2174/1574893608999140109115649.

Identification of Marker Genes for Cancer Based on Microarrays Using a Computational Biology Approach

Xiaosheng Wang
Biometric Research Branch, National Cancer Institute, National Institutes of Health, Rockville, MD 20852, U.S.A; Phone: (301)-451-5907

Xiaosheng Wang: Xiaosheng.wang@nih.gov

Abstract

Rapid advances in gene expression microarray technology have enabled to discover molecular markers used for cancer diagnosis, prognosis, and prediction. One computational challenge with using microarray data analysis to create cancer classifiers is how to effectively deal with microarray data which are composed of high-dimensional attributes (p) and low-dimensional instances (n). Gene selection and classifier construction are two key issues concerned with this topics. In this article, we reviewed major methods for computational identification of cancer marker genes. We concluded that simple methods should be preferred to complicated ones for their interpretability and applicability.

Keywords

Marker genes; Cancer; Microarrays; Computational biology

1 Introduction

Recent advances in microarray technology have made it feasible to rapidly measure the expression levels of tens of thousands of genes in a single experiment at a reasonable expense [1]. By measuring gene expression levels related to normal and tumor samples, investigators can discover molecular markers to be used for cancer diagnosis, prognosis, and prediction. Since the pioneering work of Golub et al. in applying gene expression monitoring by DNA microarray to cancer classification [2], the use of microarray technology to identify marker genes for cancer has been a hot topics in both computational and biomedical science [2–8].

Microarray data are concerned with two major issues. First, they contain a large amount of noise in gene expression data measured. Second, compared with the measured quantities of gene expression levels in experiments, the numbers of samples are severely limited. These issues bring about serious challenges for accurate identification of marker genes for cancer diagnosis and prediction. To address these issues, a substantial number of data process strategies have been investigated. These strategies are generally concerned with data normalization, feature selection and classifier construction. Actually, so many strategies have emerged that one often feels dazzled when tries to make a proper choice among them. Although there is no a unified standard in evaluation of classification methods, some basic criteria are recognized which are based on computational cost, classification accuracy and acceptance of classification models in medical applications.

2 Data Normalization

Data normalization is used to remove systematic variation in microarray experiments that may hamper proper comparisons of gene expression levels. This step is crucial to identification of marker genes as it seriously affects the subsequent analysis results. An excellent review of microarray data normalization has been given by Quackenbush [9]. The often-used normalization methods include global normalization using the global median of log intensity ratios, intensity dependent linear normalization, intensity dependent nonlinear normalization using a LOWESS curve etc [10]. In [10], the authors suggest that intensity-dependent normalization performs better than global normalization methods, and that linear and nonlinear normalization methods perform similarly by analysis of 36 cDNA microarrays of 3,840 genes obtained in an experiment to search for changes in gene expression profiles during neuronal differentiation of cortical stem cells. Dual-channel data is normalized within each array, whereas single-channel data is normalized relative to a designated reference array. There are many software tools which provide microarray data normalization methods. For example, in BRB-ArrayTools, there are four normalization methods: median normalization, housekeeping gene normalization, lowess normalization and print-tip group normalization, among which the median normalization and housekeeping gene normalization options are available for both single-channel and dual-channel data while the lowess normalization and print-tip group normalization options are available only for dual-channel data. The software can be freely downloaded from the website: http://linus.nci.nih.gov/BRB-ArrayTools.html.

3 Feature Selection

Feature selection, i.e., gene selection in microarray data, is an important step for identification of marker genes. Because the number of genes is large in a microarray data, it is tricky to select proper genes for cancer classification.

3.1 Feature Select Methods

In machine learning and data mining, the often-used feature selection methods include t-statistics, Wilcoxon-Mann-Whitney (WMW) statistics, chi-square, information gain (or information entropy) and Relief-F method etc.

The *t*-statistics and WMW statistics are two types of simple feature selection methods. The *t*-statistics measure was first used by Golub *et al*. to measure the class predictability of genes for two-class problems [2, 11]. Both *t*-statistics and WMW-statistics were used for gene selection by Dudoit et al and showed good classification performance [12].

The chi-square (χ^2) method evaluates features individually by measuring their chi-squared statistic with respect to the classes [13]. The χ^2 value of an attribute *a* is defined as follows:

$$\chi^2(a)=\sum_{v\in V}\sum_{i=1}^{n}\frac{[A_i(a=v)-E_i(a=v)]^2}{E_i(a=v)},$$

where V is the set of possible values for a, n the number of classes, $A_i(a = v)$ the number of samples in the ith class with $a = v$, and $E_i(a = v)$ the expected value of $A_i(a = v)$; $E_i(a = v) = P(a = v)P(c_i)N$, where $P(a = v)$ is the probability of $a = v$, $P(c_i)$ the probability of one sample labeled with the ith class, and N the total number of samples.

Information Gain [14] method selects the attribute with highest information gain, which measures the difference between the prior uncertainty and expected posterior uncertainty caused by attributes. The information gain by branching on an attribute *a* is defined as:

$$Info\_Gain(S,a)=E(S)-\sum_{i=1}^{n}\frac{S_i}{S}E(S_i),$$

where *E(S)* is the entropy before split, $\sum_{i=1}^{n}\frac{S_i}{S}E(S_i)$ the weighted entropy after split, and $\{S_1, S_2, \ldots, S_n\}$ the partition of sample set *S* by *a* values.

Symmetric uncertainty method compensates for information gain's bias towards features with more values. It is defined as:

$$SU(X,Y)=2\frac{IG(X|Y)}{H(X)+H(Y)},$$

where $H(X)$ and $H(Y)$ are the entropy of attribute *X* and *Y* respectively, and $IG(X \mid Y) = H(X) - H(X \mid Y)$ ($H(X \mid Y)$ is the conditional entropy of *X* given *Y*), represents additional information about *X* provided by attribute *Y*. The entropy and conditional entropy are respectively defined as:

$$H(X)=-\sum_{i}P(x_i)\log_2(P(x_i)),$$

$$H(X|Y)=-\sum_{j}P(y_j)\sum_{i}P(x_i|y_j)\log_2(P(x_i|y_j)).$$

The values of symmetric uncertainty lie between 0 and 1. A value of 1 indicates that knowing the values of either attribute completely predicts the values of the other; a value of 0 indicates that *X* and *Y* are independent.

Relief-F method estimates the quality of features according to how well their values distinguish between examples that are near to each other. Specifically, it tries to find a good estimate of the following probability to assign as the weight for each feature *a* [15]: $w_a = P$(different value of *a* | different class) − *P*(different value of *a* | same class). Differing from the majority of the heuristic measures for estimating the quality of the attributes assume the conditional independence of the attributes and are therefore less appropriate in problems which possibly involve much feature interaction. Relief algorithms (including Relief-F) do not make this assumption and therefore are efficient in estimating the quality of attributes in problems with strong dependencies between attributes [16].

In [17], the authors developed a feature selection method based on a soft-computing approach. The α depended degree was defined and utilized as the basis for gene selection. The α *depended degree* of an attribute subset *P* by the decision attribute *D* was defined as $\gamma_P(D,\alpha)=\frac{|\mathrm{POS}_P(D,\alpha)|}{|U|}$, where $0 \le \alpha \le 1$, $|\mathrm{POS}_P(D,\alpha)|=|\bigcup_{X\in U/R(D)} pos(P,X,\alpha)|$ and $pos(P,X,\alpha) = \cup \{Y \in U/R(P) \mid |Y \cap X|/|Y| \ge \alpha;\}$. When α equals to 1, the α depended degree contracts to the depended degree, an essential concept in rough set theory [18]. In [19], the authors compared the proposed feature selection method with the established methods: the depended degree, chi-square, information gain, Relief-F and symmetric uncertainty, and showed that the method was superior or comparable to the compared methods.

3.2 Wrapper vs. Filter

In the wrapper approach, the feature selection algorithm exists as a wrapper around the induction algorithm. In the other words, the feature selection algorithm searches for a good feature subset using the induction algorithm itself as part of the function evaluating feature subsets [20]. In contrast, the filter method selects features independently of any induction algorithm. In the other worlds, the filter method ignores the effects of the selected feature subset on the performance of the induction algorithm. As a result, the filter method is much faster than the wrapper method. Because microarray data contain a huge number of features (genes), the filter method is more suitable for microarray data [21].

3.3 Univariate vs. Multivariate

The univariate gene selection method evaluates the importance of each gene individually, while the multivariate gene selection method evaluates the importance of a group of genes. Obviously, the multivariate gene selection method is much more complicated than univariate gene selection method in that the former involves combinatorial searches through the space of possible feature subsets [22]. Due to a large number of genes contained in microarray data, only simplified multivariate gene selection methods are feasible [23–31]. Although the univariate feature selection approach is simple compared to the complex multivariate feature selection approaches, the former often outperformed the latter [12, 22, 32].

3.4 Number of Genes vs. Classification Performance

Although a large literature on the development and validation of predictive classifiers has emerged, most of the classifiers developed have involved complex models containing numerous genes [5, 33–38]. This has limited the interpretability of the classifiers and therefore hampered their applicability as diagnostic tools. Actually, many studies have revealed that classifiers could be developed containing few genes that provided classification accuracy comparable to that achieved by more complex models, e.g., in [3, 24, 31, 39–41], the authors explored the use of one or two genes to perform tumor classifications. They reported that the classification performance based on the one or two genes was often comparable to those based on many genes. For example, Table 1 shows that the single gene and two-gene classifiers have comparable performance to more complex classifiers in most cases examined [40–41]. It should be noted that the DLDA, k-NN, SVM and RF used a large number of genes for constructing the classifiers in most of the eleven datasets (see Table 2 in [40]).

4 Construction of Classification Rules

Many different classification rules have been proposed for high dimensional predictive classification including Support Vector Machines (SVM), Diagonal Linear Discriminant Analysis (DLDA), Artificial Neural Network (ANN), Bayesian, *k*-Nearest Neighbor (*k*-NN), Nearest Centroid (NC), Decision Tree (DT), Random Forest (RF), Rough Set (RS) [42], Emerging Pattern (EP) [43] etc. Among these classifiers, SVM, DA, ANN, GA, NB and *k*-NN produce “black-box” models, in which class predication is often based on abstract mathematical formulae which are difficult to interpret. In contrast, DT, RS and EP produce “white-box” models, which often implement classification by giving explicit rules. The “white-box” models have an advantage over the “black-box” models when applied to identification of marker genes for cancer based on microarrays for they are more understandable so as to be easily accepted by biologists and clinicians.

4.1 "Black-box" models

An SVM views input data as two sets of vectors in an *n*-dimensional space, and constructs a separating hyperplane in that space, one which maximizes the margin between the two data sets. The SVM method has been widely used in molecular classification of cancer [35, 51–53].

The Bayesian classifier is a probabilistic algorithm based on Bayes' rule and the simple assumption that the feature values are conditionally independent given the class. Given a new sample observation, the classifier assigns it to the class with the maximum conditional probability estimate. Many investigators have used the Bayesian classifier to analyze gene expression [54–57].

k-NN is an instance-based classifier. The classifier decides the class label of a new testing sample by the majority class of its *k* closest neighbors based on their Euclidean distance. Compared with SVM and Bayesian classifiers, *k*-NN is simpler while has comparable performance in classification of cancer based on gene expression data [12]. ANN has also been used for classification of cancer based on gene expression data [58–59]. Although ANN has been widely applied in biomedical fields [60–63], its utility in gene expression data is relatively unpopular due to complex of the method.

"White-box" models

DT is the rule-based classifier with non-leaf nodes representing selected attributes and leaf nodes showing classification outcomes. Every path from the root to a leaf node reflects a classification rule [14]. Some investigators have applied the method to cancer-related gene expression data [38, 64].

Rough sets is a data-analysis method originally proposed by Pawlak in the early 1980s [18], has evolved into a widely accepted machine-learning and data-mining method [42]. In [17, 39, 65–68], rough sets method was applied for cancer classification and prediction based on gene expression profiling.

The EP model developed by Li and Wong was also a "White-box" model by which they implemented classification by giving "IF-THEN"-like rules [43, 69–70]. This type of classification rules was simple, clear and efficient.

In [39, 41, 71], the authors simply constructed the classification rule based on cut-points for the expression levels of a single gene or gene pairs selected. For example, if a single gene g is selected and the expression level of the gene in the sample s is no more than T, then the sample is assigned to the class c_1; otherwise the sample is assigned to the class c_2, i.e., "$E(g, s) \leq T \Rightarrow C(s)=c_1$; $E(g, s) > T \Rightarrow C(s)=c_2$"; or a direction-reversed classification is produced, i.e., "$E(g, s) \leq T \Rightarrow C(s)=c_2$; $E(g, s) > T \Rightarrow C(s)=c_1$". Here T is the optimal cut point for gene g. The authors found the optimal cut point by using the entropy-based discretization method [72]. Obviously, this type of classification rules is simple, explicit and may be more suitable for clinical application.

Concluding Remarks

Expression profiling of marker genes for cancer can be used to develop classifiers of prognosis or sensitivity to particular treatments. However, one serious drawback of most existing methods for identification of cancer-related genes based on microarrays is that too many genes are ultimately selected for the classification of cancer, thereby hampering the interpretability of the models. Moreover, classification models based on numerous genes can also be more difficult to transfer to clinical application. Actually, it is often difficult to identify marker genes for cancer when a large cluster of genes are used to build classifiers because it is not easy to gauge which gene is essential in determining a cancerous class. In fact, some classifiers composed of very few genes can perform well. For example, Geman et al. developed the top-scoring pair(s) (*TSP*) classifier which classified gene expression profiles using a comparison-based approach [31]. The *TSP* classifier had better or comparable performance relative to multi-gene classifiers and has gained popularity [64, 73–77].

Classifier rules are often classified into two categories: “black-box” and “white-box” models. Compared with the “black-box” models, the “white-box” models are clearer, simpler and equally or even more efficient, and therefore are more inclined to be accepted in clinical applications.

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20; 270(5235):467–70. [PubMed: 7569999]
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15; 286(5439):531–7. [PubMed: 10521349]
3. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. 2002 Sep 1; 62(17):4963–7. [PubMed: 12208747]
4. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Richards WG, et al. Using gene expression ratios to predict outcome among patients with mesothelioma. J Natl Cancer Inst. 2003 Apr 16; 95(8):598–605. [PubMed: 12697852]
5. van ‘t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530–6. [PubMed: 11823860]
6. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer cell. 2002; 1(2):203–9. [PubMed: 12086878]
7. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature. 2002; 415(6870):436–42. [PubMed: 11807556]
8. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine. 2002; 8(1):68–74.
9. Quackenbush J. Microarray data normalization and transformation. Nat Genet. 2002; 32(Suppl): 496–501. [PubMed: 12454644]
10. Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R. Evaluation of normalization methods for microarray data. BMC Bioinformatics. 2003 Sep 2.4:33. [PubMed: 12950995]
11. Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics. 2004; 20(15):2429–37. [PubMed: 15087314]
12. Dudoit, S.; Fridlyand, J. Classification in microarray experiments. In: Speed, T., editor. Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC; 2003. p. 93-158.
13. Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform. 2002; 13:51–60. [PubMed: 14571374]
14. Quinlan J. Induction of decision trees. Machine Learning. 1986; 1:81–106.

15. Wang Y, Makedon FS, Ford JC, Pearlman J. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics. 2005 Apr 15; 21(8):1530–7. [PubMed: 15585531]
16. Robnik-Sikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning Journal. 2003; 53:23–69.
17. Wang X, Gotoh O. Microarray-Based Cancer Prediction Using Soft Computing Approach. Cancer Informatics. 2009; 7:123–39. [PubMed: 19718448]
18. Pawlak Z. Rough sets. International Journal of Computer and Information Sciences. 1982; 11:341–56.
19. Wang X, Gotoh O. A robust gene selection method for microarray-based cancer classification. Cancer Inform. 2010; 9:15–30. [PubMed: 20234770]
20. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997 Dec; 97(1–2):273–324.
21. Hall, MA.; Smith, LA. Practical feature subset selection for machine learning. In: McDonald, C., editor. Computer Science’98 Proceedings of the 21st Australasian Computer Science Conference ACSC’98. Perth: Springer; Feb 4–6. 1998 p. 181-91.
22. Lai C, Reinders MJT, van’t Veer LJ, Wessels LFA. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC bioinformatics. 2006; 7:235. [PubMed: 16670007]
23. Blanco R, Larranaga P, Inza I, Sierra B. Gene selection for cancer classification using wrapper approaches. Int J Pattern Recogn. 2004 Dec; 18(8):1373–90.
24. Bo T, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome biology. 2002; 3(4):RESEARCH0017. [PubMed: 11983058]
25. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002; 46(1–3):389–422.
26. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics. 2001; 17(12):1131–42. [PubMed: 11751221]
27. Silva PJS, Hashimoto RF, Kim SC, Barrera J, Brandao LO, Suh E, et al. Feature selection algorithms to find strong genes. Pattern Recogn Lett. 2005 Jul 15; 26(10):1444–53.
28. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. Genome Res. 2001; 11(11):1878–87. [PubMed: 11691853]
29. Xiong M, Li W, Zhao J, Jin L, Boerwinkle E. Feature (gene) selection in gene expression-based tumor classification. Mol Genet Metab. 2001; 73(3):239–47. [PubMed: 11461191]
30. Cong, G.; Tan, K-L.; Tung, A.; Xu, X., editors. The ACM SIGMOD International Conference on Management of Data. 2005. Mining top-k covering rule groups for gene expression data.
31. Geman D, d’Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004; 3:Article19. [PubMed: 16646797]
32. Lecocke M, Hess K. An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. Cancer Inform. 2006; 2:313–27. [PubMed: 19458774]
33. Antonov AV, Tetko IV, Mader MT, Budczies J, Mewes HW. Optimization models for cancer classification: extracting gene interaction information from microarray expression data. Bioinformatics. 2004 Mar 22; 20(5):644–52. [PubMed: 15033871]
34. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002 Jan; 18(1):39–50. [PubMed: 11836210]
35. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000 Oct; 16(10):906–14. [PubMed: 11120680]
36. Stamey TA, Warrington JA, Caldwell MC, Chen Z, Fan Z, Mahadevappa M, et al. Molecular genetic profiling of Gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia. J Urol. 2001 Dec; 166(6):2171–7. [PubMed: 11696729]

37. Li, J.; Wong, L. Advances in Web-Age Information Management. Berlin/Heidelberg: Springer; 2003. Using rules to analyse bio-medical data: a comparison between C4.5 and PCL; p. 254-65.
38. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003; 2(3 Suppl):S75–83. [PubMed: 15130820]
39. Wang X, Gotoh O. Accurate molecular classification of cancer using simple rules. BMC Med Genomics. 2009; 2:64. [PubMed: 19874631]
40. Wang X, Simon R. Microarray-based Cancer Prediction Using Single Genes. BMC Bioinformatics. 2011; 12:391. [PubMed: 21982331]
41. Wang X. Robust two-gene classifiers for cancer prediction. Genomics. 2012; 99(2):90–5. [PubMed: 22138042]
42. Pawlak, Z. Rough sets-Theoretical aspects of reasoning about data. Dordrecht ; Boston: Kluwer Academic Publishers; 1991.
43. Li J, Wong L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics. 2002 May; 18(5):725–34. [PubMed: 12050069]
44. Talantov D, Mazumder A, Yu JX, Briggs T, Jiang Y, Backus J, et al. Novel genes associated with malignant melanoma but not benign melanocytic lesions. Clinical cancer research: an official journal of the American Association for Cancer Research. 2005; 11(20):7234–42. [PubMed: 16243793]
45. Sotiriou C, Neo S-Y, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(18): 10393–8. [PubMed: 12917485]
46. Ma X-J, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer cell. 2004; 5(6): 607–16. [PubMed: 15193263]
47. Chen X, Leung SY, Yuen ST, Chu K-M, Ji J, Li R, et al. Variation in gene expression patterns in human gastric cancers. Mol Biol Cell. 2003; 14(8):3208–15. [PubMed: 12925757]
48. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98(24): 13790–5. [PubMed: 11707567]
49. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, et al. The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. N Engl J Med. 2003; 349(26):2483–94. [PubMed: 14695408]
50. Ishikawa M, Yoshida K, Yamashita Y, Ota J, Takada S, Kisanuki H, et al. Experimental trial for diagnosis of pancreatic ductal carcinoma based on gene expression profiles of pancreatic ductal cells. Cancer science. 2005; 96(7):387–93. [PubMed: 16053509]
51. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002; 46:389–422.
52. 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. [PubMed: 18647401]
53. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A. 2000; 97(1):262–7. [PubMed: 10618406]
54. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98(20):11462–7. [PubMed: 11562467]
55. Lu Y, Han JW. Cancer classification using gene expression data. Inform Syst. 2003 Jun; 28(4): 243–68.
56. Roth V, Lange T. Bayesian class discovery in microarray datasets. IEEE Trans Biomed Eng. 2004; 51(5):707–18. [PubMed: 15132496]

57. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000; 7(3–4):601–20. [PubMed: 11108481]

58. Huang CJ, Liao WC. Application of probabilistic neural networks to the class prediction of leukemia and embryonal tumor of central nervous system. Neural Processing Letters. 2004; 19:211–26.

59. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001; 7(6):673–9. [PubMed: 11385503]

60. Anagnostou T, Remzi M, Lykourinas M, Djavan B. Artificial neural networks for decision-making in urologic oncology. Eur Urol. 2003; 43(6):596–603. [PubMed: 12767358]

61. 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(3):315–29. [PubMed: 19307287]

62. Hand C. Epicenter location by analysis of interictal spikes: a case study for the use of artificial neural networks in biomedical engineering. Ann N Y Acad Sci. 2002; 980:306–14. [PubMed: 12594100]

63. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics. 2009; 10:296. [PubMed: 19765293]

64. Czajkowski M, Kretowski M. Top scoring pair decision tree for gene expression data analysis. Adv Exp Med Biol. 2011; 696:27–35. [PubMed: 21431543]

65. Sun, L.; Miao, D.; Zhang, H., editors. the 3rd International Conference on Rough Sets and Knowledge Technology. 2008. Efficient gene selection with rough sets from gene expression data.

66. Momin, BF.; Mitra, S., editors. First International Conference on Hybrid Information Technology. 2006. Reduct generation and classification of gene expression data.

67. Li, D.; Zhang, W., editors. The 1st International Conference on Rough Sets and Knowledge Technology. 2006. Gene selection using rough set theory.

68. Banerjee M, Mitra S, Banka H. Evolutinary-rough feature selection in gene expression data. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Application and Reviews. 2007; (37):622–32.

69. Li J, Liu H, Downing JR, Yeoh AE, Wong L. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics. 2003 Jan; 19(1):71–8. [PubMed: 12499295]

70. Dong, G.; Li, J., editors. the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, CA: ACM Press; 1999. Efficient mining of emerging patterns:discovering trends and differences.

71. Wang, X.; Simon, R. BMC Bioinformatics. Microarray-based Cancer Prediction Using Single Genes. in press

72. Fayyad, UM.; Irani, KB., editors. Multi-interval discretization of continuous-valued attributes for classification learning; Proceedings of the 13th International Joint Conference of Artificial Intelligence; 1993 August 28 -September 3; Chambéry: France Morgan Kaufmann;

73. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics (Oxford, England). 2005; 21(20):3896–904.

74. Leek JT. The tspair package for finding top scoring pair classifiers in R. Bioinformatics. 2009; 25(9):1203–4. [PubMed: 19276151]

75. Popovici V, Budinska E, Delorenzi M. Rgtsp: a generalized top scoring pairs package for class prediction. Bioinformatics. 2011; 27(12):1729–30. [PubMed: 21505033]

76. Edelman LB, Toia G, Geman D, Zhang W, Price ND. Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. BMC Genomics. 2009; 10:583. [PubMed: 19961616]

77. Shi P, Ray S, Zhu Q, Kon MA. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. BMC Bioinformatics. 2011; 12:375. [PubMed: 21939564]

Table 1

Comparison of classification accuracy (%)

Method Dataset	SGC-t	SGC-W	TGC-1	TGC-2	TSP	DLDA	*k*-NN	SVM	RF
Melanoma [44]	97	96	97	96	99	97	97	97	97
Breast Cancer 1 [45]	63	69	64	64	75	61	53	52	43
Brain Cancer [7]	80	77	77	75	77	65	73	60	70
Breast Cancer 2 [46]	58	50	82	78	47	73	67	73	67
Gastric Tumor [47]	89	80	89	88	91	81	96	97	95
Lung Cancer 1 [48]	98	95	98	100	95	95	98	98	98
Lung Cancer 2 [3]	93	93	93	93	97	99	99	99	99
Lymphoma [8]	74	71	59	60	57	66	52	59	57
Myeloma [49]	68	67	68	54	71	75	78	74	79
Pancreatic Cancer [50]	69	90	71	73	90	63	61	65	55
Prostate Cancer [6]	89	89	89	90	81	78	93	93	93

Note:

1 SGC-t: Single Gene Classifier with the t-test gene selection method.

2 SGC-W: Single Gene Classifier with the WMW gene selection method.

3 TGC-1: Two Gene Classifier Type 1.

4 TGC-1: Two Gene Classifier Type 2.

5 TSP: Top-Scoring Pair(s) (TSP) classifier.

6 DLDA: Diagonal Linear Discriminant Analysis.

7 *k*-NN: *k*-Nearest Neighbor (k=3).

8 SVM: Support Vector Machines.

9 RF: Random Forest.

10 Leave-one-out cross validation (LOOCV) results are presented.