BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 11 2005, pages 2691–2697
doi:10.1093/bioinformatics/bti419

Genetics and population analysis

Multiclass cancer classification and biomarker discovery using GA-based algorithms

Jane Jijun Liu[1], Gene Cutler[1], Wuxiong Li[1], Zheng Pan[1], Sihua Peng[2], Tim Hoey[1], Liangbiao Chen[3] and Xuefeng Bruce Ling[1,*]

[1]Tularik Inc., South San Francisco, CA 94080, USA, [2]Zhejiang University, Hangzhou 310027, China, and [3]Institute of Genetics and Developmental Biology, The Chinese Academy of Sciences, Beijing 100101, China

Received on August 29, 2004; revised on March 11, 2005; accepted on March 30, 2005
Advance Access publication April 6, 2005

ABSTRACT

Motivation: The development of microarray-based high-throughput gene profiling has led to the hope that this technology could provide an efficient and accurate means of diagnosing and classifying tumors, as well as predicting prognoses and effective treatments. However, the large amount of data generated by microarrays requires effective reduction of discriminant gene features into reliable sets of tumor biomarkers for such multiclass tumor discrimination. The availability of reliable sets of biomarkers, especially serum biomarkers, should have a major impact on our understanding and treatment of cancer.

Results: We have combined genetic algorithm (GA) and all paired (AP) support vector machine (SVM) methods for multiclass cancer categorization. Predictive features can be automatically determined through iterative GA/SVM, leading to very compact sets of non-redundant cancer-relevant genes with the best classification performance reported to date. Interestingly, these different classifier sets harbor only modest overlapping gene features but have similar levels of accuracy in leave-one-out cross-validations (LOOCV). Further characterization of these optimal tumor discriminant features, including the use of nearest shrunken centroids (NSC), analysis of annotations and literature text mining, reveals previously unappreciated tumor subclasses and a series of genes that could be used as cancer biomarkers. With this approach, we believe that microarray-based multiclass molecular analysis can be an effective tool for cancer biomarker discovery and subsequent molecular cancer diagnosis.

Contact: xuefeng_ling@yahoo.com

Supplementary information: http://www.fishgenome.org/publication/Liu/bioinformatics/

INTRODUCTION

Traditional cancer diagnosis relies on a complex and inexact combination of clinical and histopathological data. These classic approaches may fail when dealing with atypical tumors or morphologically indistinguishable tumor subtypes. Advances in the area of microarray-based expression analysis have led to the promise of cancer diagnosis using new molecular-based approaches; however, this new type of data presents new challenges. Microarray-based tumor comparative hybridization profiles have been introduced for binary and multiclass cancer classifications (Ross *et al.*, 2000; Ramaswamy *et al.*, 2001; Su *et al.*, 2001; Yeang *et al.*, 2001; Ooi and Tan, 2003; Peng *et al.*, 2003). The high-throughput nature of this technology makes it feasible to diagnose a large number of common malignancies in parallel based on comprehensive microarray datasets. However, multiclass tumor classification using large-scale expression databases has significant statistical and analytical implications.

The support vector machine (SVM) algorithm (Vapnik, 1998) has been one of the most powerful supervised learning algorithms in biological data analysis including microarray-based expression analysis (Brown *et al.*, 2000; Furey *et al.*, 2000), remote protein homology detection (Jaakkola *et al.*, 1999; Ben-Hur and Brutlag, 2003) and translation initiation site recognition (Zien *et al.*, 2000). Utilized as binary categorical classifiers, the SVM method has been shown to consistently outperform other classification approaches including weighted voting and k-nearest neighbors (Ramaswamy *et al.*, 2001). To extend the SVM algorithm to multiclass classification, integration with another algorithm such as the one-versus-all (OVA) or all-paired (AP) binary comparisons is required.

The genetic algorithm (GA), as introduced by Goldberg (1989), performs randomized search and optimization mimicking evolution and natural genetics involving at least three types of genetic operators: selection, cross over and mutation. These algorithm implementations benefit from parallel population-based searches in combination with stochastic genetic operations, distinguishing them from other search methods (Goldberg, 1989; Goldberg and Deb, 1991). In microarray expression data analysis, GA has been effectively employed for binary (Li *et al.*, 2001) and multiclass cancer discrimination (Ooi and Tan, 2003; Peng *et al.*, 2003).

Advances in the area of microarray-based cancer diagnosis promise to greatly advance cancer diagnosis, especially in situations where tumors are clinically atypical. In addition, microarray-based classification schemes can potentially detect previously unrecognized tumor subtypes, discover subtype-associated biomarkers and improve diagnosis accuracy. The nearest shrunken centroids method (NSC), based upon an enhancement of the simple nearest prototype (centroid) classifier, has been previously used for tumor subclass discovery and was demonstrated to be able to identify succinctly

*To whom correspondence should be addressed at Amgen San Francisco, South San Francisco, CA 94080, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Current address: Amgen Inc., South San Francisco, CA 94080, USA discriminant genes for each predicted category (Tibshirani *et al.*, 2002, 2003).

The lack of specificity of most cancer biomarkers, including the widely used prostate-specific antigen (PSA) leads to many false positives (Brawer, 2000). The prevailing view is that, even though discriminant biomarkers for most solid tumors are not available, the most powerful single cancer biomarkers may have already been discovered (Diamandis, 2004). Technological advances in the areas of high-throughput expression profiling, high-density tissue microarrays, clinical pathology and bioinformatics have raised expectations as a small, but growing, number of candidate serum biomarkers have been identified using array-based observations (Zhou *et al.*, 1998; Dhanasekaran *et al.*, 2001; Mok *et al.*, 2001; Kim *et al.*, 2002; Rubin *et al.*, 2002). With the completion of the human genome sequencing and large sets of high-quality annotations made using automated algorithms (Pouliot *et al.*, 2001; Harris *et al.*, 2004), global methods have been utilized to perform large-scale delineation of all secreted proteins in the search for biomarkers overexpressed in cancer tissue and serum (Welsh *et al.*, 2003).

As briefly discussed in our recent report (Peng *et al.*, 2003), we have combined the GA and SVM algorithms, and utilized GA to evolve AP-SVM-based multiclass cancer classifiers. This algorithm has derived compact discriminant sets of cancer-related genes with the highest leave-one-out cross-validation (LOOCV) accuracies seen in the multiclass tumor classification literature. These predictive gene sets can achieve similar LOOCV accuracies but only overlap modestly, suggesting that less sensitive or specific genes could be discovered and may be applied in panels to devise diagnostic methodologies with improved sensitivity and specificity. In addition to the GA-based efforts, we utilized the NSC approach to post-process the GA-selected optimal feature predictors in an effort to discover previously unappreciated cancer subtypes and the associated biomarkers that best characterize them. We have used gene annotation, text mining and literature searches to categorize the biological functions of the GA-selected features as well as their cellular localization and biomarker relevancy. This analysis has revealed previously validated extracellular and membrane-bound tumor biomarkers along with many novel ones. By using our innovative analytical approaches, it is likely that microarray-based multiclass molecular analysis can be utilized for tumor biomarker discovery and effective molecular cancer diagnosis.

MATERIALS AND METHODS

Datasets

The NCI60 data (Ross *et al.*, 2000) contains gene expression profiles of 9712 spotted cDNAs in 68 cancer cell lines. The unknown, normal breast, lymph node and prostate cancer cell lines, due to their small numbers, were excluded from further analysis, leaving 61 cell lines with nine sites of origin. As previously described at http://genome-www.stanford.edu/nci60/help.shtml (Ross *et al.*, 2000), following hybridization, arrays were scanned using a laser-scanning microscope. Separate images were acquired for Cy3 and Cy5. CH2D/CH1D fluorescent ratios were utilized as measures for mRNA relative abundance. There are a total of 6116 genes with standard deviations ranging from 0.32 to 2.63 (Supplementary Fig. 1). The top 1000 genes (standard deviation >0.99) were selected as the truncated dataset and were analyzed throughout this study as the 'NCI60 dataset'.

The Brown dataset (Munagala *et al.*, 2004) consists of 268 primary tumor samples analyzed on spotted cDNA arrays containing 7452 probes. The uterus cancer, carcinoid and adrenal benign adenoma tumors were excluded from further analysis due to their small numbers. The remaining 260 tumor samples span 15 tissue types: breast, central nervous system, kidney, lung, ovary, pancreas, prostate, soft tissue, stomach, bladder, liver, lymph node, skin, testis and colon. Detailed protocols and techniques are available at http://cmgm.stanford.edu/pbrown/. The primary data tables can be obtained at http://microarray-pubs.stanford.edu/margin_clus/. Tumor samples or their associated cancer classes were excluded from the subtype analysis due to either the lack of clinical subtype information or the small sample size within the cancer categories.

GA/SVM/NSC multiclass tumor classification

The GA/SVM algorithm was used as previously described (Peng *et al.*, 2003) with a generation number of 100 000, a population size of 40 and a chromosome size of 40. These steps were followed by NSC classification (Tibshirani *et al.*, 2002, 2003) to identify potential tumor subtypes. The NSC algorithm has been implemented using the 'pamr' package of the R-project. Two hundred iterations of NSC cross-validation were performed to optimize the size of selected gene sets and minimize the error rate. Genes whose shrunken class centroids reached zero for the particular tumor class were eliminated. Average-linkage hierarchical clustering of a centered Pearson correlation similarity matrix (Hastie *et al.*, 2001) was applied using log-transformed gene expression values. Implementation details, including source code, are available in the Supplementary Material.

Assembling comprehensive annotation information

The NCI60 probe gene sets have been annotated according to the Gene Ontology specifications (http://www.geneontology.org) using the GoPipe system http://gopipe.fishgenome.org/ (Chen *et al.*, 2005) and the Ingenuity Pathways Analysis application (Palo Alto, http://www.ingenuity.com). Detailed information can be found in the Supplementary Material.

RESULTS

GA/SVM method allows highly accurate multiclass tumor classifications

Our GA/SVM algorithm consists of three main components: a GA-based gene selector, SVM-based binary classifiers distinguishing between tumor samples and multiclass categorization by an AP/SVM voting strategy (Supplementary Fig. 2A). The iterative GA's genetic selection and evolution fitness test achieved substantial feature reduction, leading to compact sets of non-redundant discriminant genes. NCI60 (Ross *et al.*, 2000), a dataset of cell lines corresponding to nine tumor types, has been utilized extensively to compare various methods of classification on microarray expression analysis (Golub *et al.*, 1999; Ooi and Tan, 2003; Peng *et al.*, 2003). Before the report of our GA/SVM algorithm (Peng *et al.*, 2003; and this study), GA/MLHD, a GA approach in combination with the MLHD classification method, had been found to have the best multiclass tumor distinction accuracies on the NCI60 datasets (Ooi and Tan, 2003). However, our replication of the GA/MLHD data analysis (Ooi and Tan, 2003) revealed that Ooi *et al.* had incorrectly labeled some samples in their processed NCI60 dataset. The labeling inconsistency can be identified by examination of the header of the NCI_full_train.txt file online at the author's website http://www.omniarray.com/bioinformatics/GA/full_data/ (Supplementary Fig. 3). Contrary to the fact that there should be four samples labeled '2—central nervous system tumor' and five samples labeled '8—renal tumor' as described in Ooi *et al.*'s Supplementary Table S3a (Ooi and Tan, 2003), NCI_full_train.txt data file header shows three central nervous system tumor samples and six renal tumor samples. A set of 1000 genes had been derived (Ooi and Tan, 2003) to demonstrate the predicting performance of the GA/MLHD algorithm.

Table 1. Effects of GA parameters on two different GA-based multiclass tumor classifications[a]

Dataset	Crossover	Selection	*Pc*	*Pm*	Number of features	LOOCV (%)	Algorithm	Reference
NCI60[b] (1000)	Uniform	SUS	1	0.002	13	85.37	GA/MLHD	Ooi and Tan (2003)
NCI60 (1000)	Uniform	SUS	1	0.002	12	70.73	GA/MLHD	This study
NCI60 (1000)	Uniform	SUS	1	0.002	40	88.52	GA/SVM	This study
NCI60[b] (1000)	One-point	RWS	0.8	0.02	13	75.61	GA/MLHD	Ooi and Tan (2003)
NCI60 (1000)	One-point	RWS	0.8	0.02	12	68.29	GA/MLHD	This study
NCI60 (1000)	One-point	RWS	0.8	0.02	40	86.89	GA/SVM	This study

[a] Microarray dataset: NCI60. Two GA selection methods: stochastic universal sampling (SUS); roulette wheel selection (RWS).

[b] Datasets revealed to be inconsistently processed.

Table 2. Comparison of GA/SVM on NCI60 dataset with other previously described multiclass tumor distinction methods

Classification method	LOOCV (%)	Number of features (genes)	Reference
Hierarchical clustering	81	6831	Ross *et al*. (2000)
GA/MLHD	85.37	13	Ooi and Tan (2003)
GA/MLHD	70.73	12	This study
GA/SVM	88.52	40	This study

Since this 1000 gene set was derived from the full (erroneous) dataset, its validity is questionable as well.

The two GA-based multiclass tumor classification algorithms, GA/MLHD and GA/SVM, were compared in this study on our processed NCI60 dataset. The effects of different GA parameters and numbers of selected features on LOOCV results are summarized in Table 1. The GA/MLHD did not perform as well as was previously described—the LOOCV accuracy of 70.73% we observed was a significant drop from the reported 85.37% (Ooi and Tan, 2003). In comparison, GA/SVM achieved an LOOCV accuracy of 88.52% on the same dataset (Table 2), indicating that AP/SVM significantly outperformed the MLHD approach as a categorical classifier.

GA/SVM algorithm derives multiple compact predictor sets with similar classification accuracies

We have processed the NCI60 expression dataset for multiclass tumor discrimination. For this analysis, the GA/SVM parameters were preset with a population size of 40, a maximum number of generations of 100 000 and a chromosome size of 40. One hundred series of independent GA/SVM operations were performed with random initial seeding from the NCI60 gene set (Supplementary Fig. 2B). Upon the completion of all GA generations, the best predictive feature set from each final generation was compared with all the others to locate the set with the highest overall accuracy. The 100 generated predictor sets fall into a narrow range of LOOCV accuracies, between 78.69 and 88.52% (Fig. 1A). Thus, these optimal sets evolving from different randomly seeded populations yield reproducibly accurate multiclass tumor classifications.

After sorting the selected genes by the standard deviation of their expression values, the frequency of selection of gene features across all the optimal predictive set series was plotted (Fig. 1B). Interestingly, a small number of the genes have been repeatedly selected beyond what would be expected at random. Details of the genes and their selection frequencies are summarized in the Supplementary Material. The optimal predictive sets of 40 features each harbor modest overlapping genes. The numbers of pairwise overlaps range from 0 to 12 with a mode of 4 (Supplementary Fig. 4A). For comparison, a set of 100 randomly selected gene sets have pairwise overlaps in the range of 0–7 with a mode of 2. Only 208 genes of the total 1000 NCI60 genes were never selected. Among the 792 selected genes, most were selected no more than three times, but there exists a small group of frequently selected genes in a pattern not seen with the random gene sets (Supplementary Fig. 4B). This observation suggests that the GA-based feature selection approach can lead to the identification of largely independent optimal predictor gene sets.

Characterization of GA/SVM-derived optimal predictors—feature annotation and text mining for cancer biomarker discovery

The gene index, GA selection frequency, gene name and description for each predictor have been summarized and shown in the Supplementary Material. The genes of the optimal predictor sets have been annotated with the GO ontology (http://www.geneontology.org) using GoPipe (http://gopipe.fishgenome.org/, Chen *et al*., the Chinese Academy of Sciences, unpublished) and the Ingenuity Pathways Analysis tool (Palo Alto, http://www.ingenuity.com). In general, the annotation efficiency for the set of the gene predictors is low. Among the total of 792 GA-derived NCI60 gene features, only 39.9% have biological localization annotation, and only 23.6% have defined biological functions (Supplementary Fig. 6A and B).

The GA/SVM-selected genes were further annotated with the Ingenuity Pathways Analysis tool for biological pathway analysis. The NCI60 optimal gene features were clustered into 56 interaction networks (Supplementary Tables 1 and 2, and Supplementary Material of pathway network diagrams). The most common pathways in the Ingenuity networks were ERK/MAPK signaling, Wnt/beta-catenin signaling, p38 MAPK signaling, cell cycle G1/S checkpoint regulation, PI3K/AKT signaling, apoptosis signaling, cell cycle G2/M DNA damage checkpoint regulation, B cell receptor signaling, insulin receptor signaling, SAPK/JNK signaling and TGF-beta signaling. Most of the biological pathways revealed in this study are already known to be involved in tumorigenesis. In line with our observation, the Wnt/beta-catenin pathway, an important pathway in tumorigenesis, has been previously identified through microarray-based multiclass cancer analysis (Ramaswamy *et al*.,

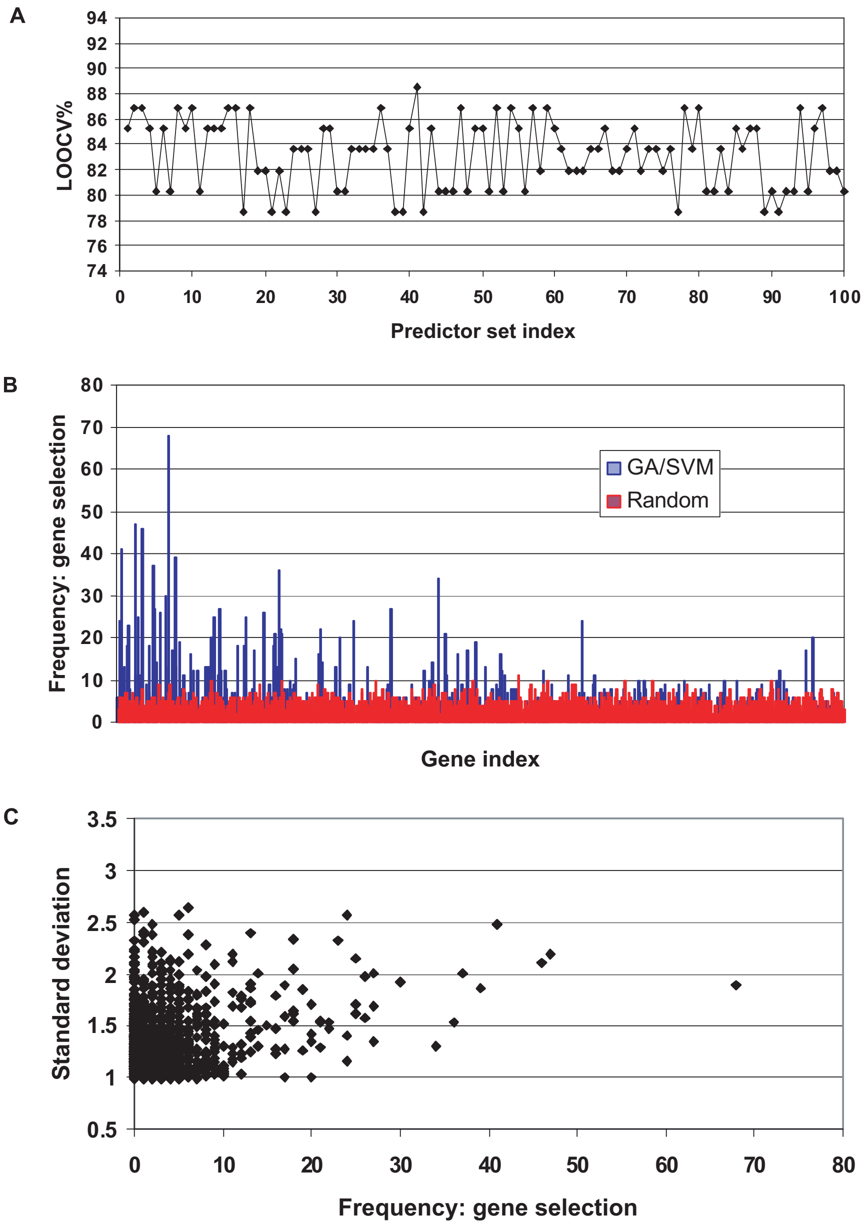

Fig. 1. Analysis of the GA/SVM-derived NCI60 optimal feature sets. (**A**) LOOCV accuracy is plotted for each predictor set generated by GA/SVM. (**B**) Frequency of gene selection into the 100 optimal predictor sets (blue). Genes have been sorted from high to low standard deviations of expression (left to right). Frequencies of selection of genes in 100 randomly generated sets are shown for comparison (red).

2001). Deregulation of these pathways alters cell proliferation, cell survival, adhesion and migration, all of which can potentially lead to tumorigenesis.

Unfortunately, a number of important cancer biomarkers including prostate specific antigen (PSA), CA125, CA15.3, CA19.9, MIC-1, alpha fetoprotein (AFP), kallikreins 6 and 10, prostasin, human chorionic gondaotropin (hCG) and bombesin are not represented in the NCI60 dataset and therefore would not be discovered by our microarray-driven approach. To address the potential clinical utility of the GA/SVM-derived optimal genes for cancer biomarker studies, the selected genes were annotated and those with protein products localizing to the extracellular space or plasma membrane were selected and analysed. A heat map representation using these antigen biomarkers was generated (Fig. 2). In the heat map, color is proportional to the log-transformed expression values of the genes, with red representing high expression and green low expression.

GA/SVM-derived predictors were further characterized through the NSC method, revealing previously unappreciated tumor subtypes

Classification using NSC was performed on our GA/SVM-derived predictors to identify potential subtypes of the nine cancer classes in our NCI60 dataset. This was done to reveal potential tumor subclasses and their associated biomarkers. It has been shown

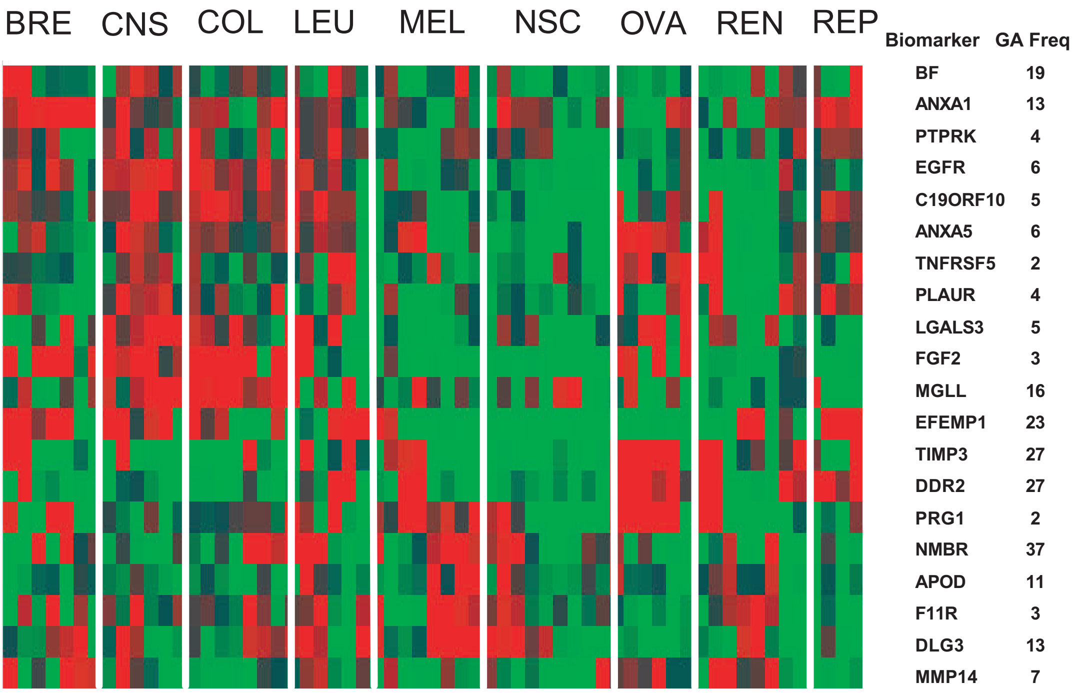

Fig. 2. Potential biomarker panel expression heat map. Rows represent genes and columns represent NCI60 cancer cell lines, ordered by hierarchical clustering in the gene dimension. Red represents high expression levels and green represents low expression levels. Tissue of origin of the cell lines: breast (BRE), central nervous system (CNS), colon (COL), leukemia (LEU), melanoma (MEL), non-small-cell-lung (NSC), ovarian (OVA), renal (REN) and reproductive (REP).

previously (Tibshirani *et al.*, 2002, 2003) that the 'shrunken' centroids generated by NSC can be utilized to identify the subsets of genes that best characterize each category. The NSC classification computes the standard nearest centroid for each class and 'shrinks' each class centroid toward the overall centroid for all classes. This shrinkage consists of moving the centroid towards zero sequentially by predefined intervals, with the amount of optimal shrinkage determined by cross validation. New samples are then assigned to the nearest shrunken class centroid.

In order to find a balance between classification accuracy and gene marker numbers, 200 iterations of NSC CVs were performed. The NSC model with the fewest genes contained 678 genes, had a shrinkage amount of 0.72 and an error rate of up to 32.8%. The best NSC model achieved contained 779 genes, had a shrinkage amount of 0.83 and an error rate of 24.6%. In comparison, a previous NSC classification with zero subclasses using the NCI60 dataset yielded an error rate of 42.8% (Tibshirani *et al.*, 2003). The NSC-selected subsets of genes for each specific tumor type were utilized for tumor subtype characterization and potential discovery of novel cancer subtypes. Divisions of the tumor types into putative subtypes based on gene expression through manual examination of heat maps generated for each tumor type are shown in Supplementary Figure 7. Potential subclass patterns can be found for renal cancer, colon cancer, breast cancer and melanoma.

Although NCI60 cancer cell lines have been extensively used as experimental models of neoplastic diseases, such cell lines are fundamentally different from both normal and cancerous tissues. Thus, we have applied our GA based algorithm to a more comprehensive tumor microarray dataset, the Brown data (Munagala *et al.*, 2004). This dataset contains gene expression profiles for 260 primary tumor samples spanning 15 different classes. Consistent with the high performance of multiclass cancer discrimination seen with the NCI60, the GA/SVM algorithm achieved 81.23% LOOCV accuracy with the Brown dataset, and derived a combined set of 730 optimal gene features from 100 GA classifier sets of similar LOOCV accuracies. Further NSC analysis was utilized to select gene markers that best characterized each tumor subtype. As shown in Figure 3, unsupervised hierarchical clustering of the NSC-selected gene features' expression data readily separated and largely agreed with clinically observed tumor subtypes in CNS, kidney, ovary and soft tissue tumor classes. In addition, the expression heat map (Fig. 3) revealed distinguishable patterns within the clinically observed tumor subtypes. The RNA-level heterogeneity among tumor subtype samples, reflected by this GA based analysis, may expose previously undetected tumor heterogeneity. In both the NCI60 tumor cell line and Brown human tumor datasets, our GA based molecular analysis discovered potentially novel, clinically unappreciated tumor categories.

DISCUSSION

As seen in our array-based cancer classification study, microarray technology is a powerful taxonomic tool because of its ability to discriminate between very different cell types. However, our use of array datasets to model cancer classifiers and to derive biomarkers with diagnostic potential does not lead to the determination of which genes and networks are causative in complex diseases like human cancers. Microarray hybridization signals are not necessarily linearly related to true gene expression levels; the relationship being complex and dependent on many factors. Thus, one should be cautious in using microarray-prioritized genes in the molecular understanding

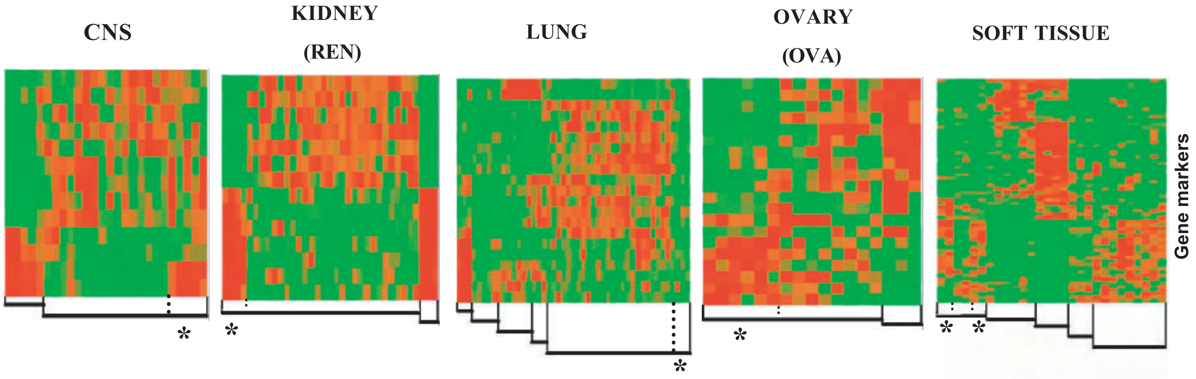

Fig. 3. NSC cancer subtype analysis in CNS, kidney, lung, ovary and soft tissue tumor classes. From left to right tumor panels, each solid box represents a particular tumor subclass in the Brown dataset: central nervous system (CNS), subclasses of medulloblastoma and glioblastoma; kidney, subclasses of renal cell carcinoma (RCC) clear cell, RCC chromophobe; lung, subclasses of small cell, large cell, squamous, unknown, adeno; ovary, subclasses of serous papillary carcinoma, poorly/undifferentiated adenocarcinoma; soft tissue, subclasses of malignant fibrous histocytoma (MFH), gastrointestinal stromal tumor (GIST), benign solitary fibrous tumor (SFT), liposarcoma, leiomyosarcoma. Asterisks indicate potential novel or unappreciated subclasses.

of tumor development and behavior, an area that requires biological validation and extensive bioinformatics (Miklos and Maleszka, 2004).

We report here our novel analytical methods for multiclass cancer classification and gene biomarker discovery. Our GA/SVM method has achieved the best LOOCV accuracy performances of all reported methods, both for 9-class (NCI60) and 14-class (GCM, Online Supplementary Material) cancer classifications. The GA-based gene selection leads to significant feature reduction. For example, with this methodology, 40 genes are sufficient to allow highly accurate multiclass tumor distinctions. This is in contrast to the previously described SVM-based classification method (Ramaswamy *et al.*, 2001) in which all 16063 GCM genes were required to ensure optimal classification performance. Any feature reductions in that context compromised predictive accuracies. Our AP/SVM categorical classifier significantly outperformed the previously reported MLHD approach (Ooi and Tan, 2003). This observation is consistent with previous observations (Furey *et al.*, 2000) that SVM is the system of choice for expression studies where data can be sparse and noisy, and each experiment usually contains thousands of expression measurements. The significant feature reduction capability and high classification performance of the GA/SVM method offer a practical solution for microarray-based clinical diagnosis of tumors.

A unique feature of the GA-based classification scheme is that a random initial seeding of the GA gene selection can lead to independent optimal predictive sets, which are compact, and with similar classification accuracies, yet with only modest overlaps in gene content. Although each of the selected gene sets performs with high accuracy, none of the component genes alone offers the adequate sensitivity, specificity or predictive value needed for multiclass tumor distinctions. Therefore, our GA-based gene selection method is capable of discovering features which are less individually sensitive or specific but which could be used in small panels with robust classification performance. Most of the current established biomarkers, such as PSA, are single genes that generally are not robust enough to give accurate predictions for large-scale population screening (Brawer, 2000). Though not yet validated in expression-based cancer diagnosis, the application of biomarker panels for cancer diagnosis using methodologies including mass spectrometry and biochip-based screening has been under active investigation (Petricoin *et al.*, 2002). The hypothesis is that biological molecules can work in panels for cancer diagnosis as long as the relative abundance in the panels of ‘informative molecules’ is sufficient to discriminate and reflect ongoing physiological and pathological events during tumorigenesis. We propose that the same premise can also be applied to microarray-based cancer diagnosis and biomarker discovery platforms. The GA-based process described here may allow for the discovery of biomarkers which might otherwise be missed due to their individually low sensitivity or specificity, but which in panels could be used for diagnostic approaches with improved sensitivity and specificity.

As shown in Supplementary Figure 5, the modest overlaps between the gene sets cannot be simply explained by a comparison of feature selection frequency and/or feature expression variance. As pointed out previously (Ooi and Tan, 2003), the GA-based approach avoids reliance upon a rank-based gene selection scheme, which tends to select genes that share similar hybridization patterns across the dataset. Thus, the multiclass cancer distinction feature panels derived from GA/SVM may turn out to be more useful for biological discovery and understanding of the biology underlying tumorigenesis. The extreme complexity and redundancy of the gene expression pathways in different cancer types are likely to have contributed to what we have observed, with different gene panels having similar accuracies in cancer classification. It is also likely that some putative biomarkers that would be strongly selected by our approach are simply missing from the starting microarray gene set. Further application of the GA/SVM algorithm to larger and more comprehensive datasets, such as those derived from whole genome arrays as well as larger tumor sets, will shed new light on this issue.

Classical cancer diagnosis has been largely based upon the subjective interpretation and examination of a tumor’s morphology and depends on highly skilled pathologists. However, many tumors are morphologically ambiguous and hard to classify using this approach. Our array-based GA/SVM approach offers the promise of objective and accurate tumor classification. Subsequent NSC cancer subtyping and biomarker selection, using the NCI60 and Brown cancer datasets, demonstrate previously unappreciated sub-categories where heterogeneous gene expression occurs within morphologically indistinguishable cancer samples. Our GA based algorithm may facilitate the discovery of novel surrogate markers and clinically relevant molecular subtypes, providing clinicians with new information for improved cancer diagnosis, prognosis and treatment stratification.

Ideally, tumor biomarkers would be assayed non-invasively and economically in blood or other readily accessible biological samples. Thus, the best cancer biomarkers would be secreted (Welsh *et al.*, 2003) or otherwise shed into the circulatory system during tumorigenesis. To determine how our selected genes matched these criteria, we annotated them for the expected localization of their protein products. We compiled a set of 20 genes from our larger GA selected NCI60 gene set (Supplementary Table 3) that we predict to be optimal biomarkers, with protein products localized in the extracellular space or cell membrane. As shown in the heat map in Figure 2, this 'optimal' antigen panel gives an LOOCV accuracy of 82% for array-based multiclass tumor distinction. Additional histochemical analyses will be needed to further validate the clinical usefulness of these biomarkers. Nevertheless, this result has significant clinical implications: molecular tumor classification can be done relatively non-invasively in the clinic using a relatively small set of antigen biomarkers. We believe that this approach can lead to quicker, more accurate, less invasive and more comprehensive tumor identification. This should have tremendous implications for the future of oncology.

ACKNOWLEDGEMENTS

We are thankful to Tularik scientists Drs. Kenneth Luskey, Scott Powers, Paul Kassner, Steve Young and Kang Dai for critical discussions, and to the authors of previous publication who made their source code available on the Internet. We are also thankful to IT groups at both the Chinese Academy of Sciences (Sheng Zhu) and Tularik Inc. (Denise Stockman, Juliette Page) for their outstanding support.

REFERENCES

Ben-Hur,A. and Brutlag,D. (2003) Remote homology detection: a motif based approach. *Bioinformatics*, **19** (Suppl. 1), i26–i33.

Brawer,M.K. (2000) Prostate-specific antigen. *Semin. Surg. Oncol.*, **18**, 3–9.

Brown,M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Chen,Z. *et al.* (2005) GoPipe: streamlined Gene Ontology annotation for batch anonymous sequences with statistics. *Prog. Biochem. Biophys.*, **32**, 187–191.

Dhanasekaran,S.M. *et al.* (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.

Diamandis,E.P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell Proteom.*, **3**, 367–378.

Furey,T.S. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Goldberg,D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, New York.

Goldberg,D.E. and Deb,K. (1991) A comparative analysis of selection schemes used in genetic algorithms. In Rawlins,G. (ed.), *Foundations of Genetic Algorithms*. Morgan Kaufman, Berlin, pp. 69–93.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32** (Database issue), D258–D261.

Hastie,T. Tibshirani,R. and Friedman,J. (2001) *The Elements of Statisitcal Learning Data Mining, Inference, and Prediction*. Springer, New York, pp. 453–480.

Jaakkola,T. *et al.* (1999) Using the Fisher kernel method to detect remote protein homologies. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 149–158.

Kim,J.H. *et al.* (2002) Osteopontin as a potential diagnostic biomarker for ovarian cancer. *JAMA*, **287**, 1671–1679.

Li,L. *et al.* (2001) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Comb. Chem. High Throughput Screen*, **4**, 727–739.

Miklos,G.L. and Maleszka,R. (2004) Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.*, **22**, 615–621.

Mok,S.C. *et al.* (2001) Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology. *J. Natl Cancer Inst.*, **93**, 1458–1464.

Munagala,K. *et al.* (2004) Cancer characterization and feature set extraction by discriminative margin clustering. *BMC Bioinform.*, **5**, 21.

Ooi,C.H. and Tan,P. (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, **19**, 37–44.

Peng,S. *et al.* (2003) Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Lett.*, **555**, 358–362.

Petricoin,E.F. *et al.* (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577.

Pouliot,Y. *et al.* (2001) DIAN: a novel algorithm for genome ontological classification. *Genome Res.*, **11**, 1766–1779.

Ramaswamy,S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Ross,D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.

Rubin,M.A. *et al.* (2002) alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *JAMA*, **287**, 1662–1670.

Su,A.I. *et al.* (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Biol.*, **61**, 7388–7393.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tibshirani,R. *et al.* (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, 104–117.

Vapnik,V. (1998). *Statistical Learning Theory*. Wiley.

Welsh,J.B. *et al.* (2003) Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. *Proc. Natl Acad. Sci. USA*, **100**, 3410–3415.

Yeang,C.H. *et al.* (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **17** (Suppl. 1), S316–S322.

Zhou,W. *et al.* (1998) Identifying markers for pancreatic cancer by gene expression analysis. *Cancer Epidemiol. Biomarkers Prev.*, **7**, 109–112.

Zien,A. *et al.* (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.