

# Higher Dimensional Approach for Classification of Lung Cancer Microarray Data

F. Crimins R. Dimitri T. Klein  
N. Palmer L. Cowen \*  
Department of Computer Science  
Tufts University  
Medford, MA 02155

## ABSTRACT

A lung cancer microarray dataset is re-examined using simple techniques, but retaining more of the high-dimensional structure. Genes of potential biological importance are also uncovered.

## 1. INTRODUCTION

The CAMDA 2003 competition involves the re-analysis of lung cancer microarray datasets. Four separate studies sought clusters that correlated with survival among patients diagnosed with adenocarcinoma, the most common type of lung cancer. Two of these studies, the one of Bhattacharjee et al. [1] and the one of Garber et al. [5] included samples not only from adenocarcinoma, but also microarray data from several other types of lung cancer tumors, as well as normal lung tissue. A secondary goal of these two papers was to construct a classifier that could distinguish between expression data for each of the different lung-cancer types plus normal lung tissue, and, in addition, find a small set of expression vectors that could account for the difference. It is this easier classification problem that is the subject of the present paper.

Bhattacharjee et al. obtained a dataset of 186 patients with four clinically distinct types of lung cancer plus normal lung tissue. The data consists of 12,600 transcript sequences for each patient. Garber et al. looked at five types of lung cancer (three of which were equivalent to three of the four considered by [1]) plus normal lung tissue over 71 patient samples (from 56 patients) over 23,100 transcript sequences (representing 17,108 unique genes). Because the number of transcript sequences was very large, both groups of researchers first identified a subset of transcript sequences that had “meaningful” expression data. In particular, [1] identified a subset of 3,312 “most variable” genes over the five different lung tissue types, and then used a subset of 675 of these to construct

\*To whom correspondence should be addressed: cowen@cs.tufts.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CAMDA '03, November 2003, City, State.

subclusters of the adenocarcinoma subtype. [5] searched transcript sequences that were similar among tumor pairs, but varied most widely among all tumor samples, yielding a subset of 918 (representing 835 unique genes).

In this abstract, we show that there is something to be gained by studying the entire dataset without first doing such preliminary dimension reduction—that such an approach can yield better results than immediately restricting to a small subset of transcript vectors whose values, considered individually, appear to contain the most discriminatory information. In this sense, our work supports the study of Li et al. [8] in CAMDA 2000, that also cautioned against looking at single expression vectors in isolation, in order to determine their discriminatory power.

While processing 12,600-dimensional data is beyond the scope of most commercial software packages designed for these problems, it is only a minor headache to code simple non-parametric classifiers that can handle the full dimensionality of the dataset in C++. First we show that perhaps the simplest non-parametric classifier, k-nearest-neighbor, performs extremely well on the five-class problem considered by [1] in cross-validation. That this is such an easy problem is perhaps not too surprising, given that these classes of lung tissue are also clinically identifiable as distinct. We then go on to the more interesting question of finding a small set of transcript sequences that can by themselves discriminate between the five classes. This is important because small sets of transcript sequences that distinguish between the classes may correspond to genes that are biologically important toward understanding the underlying lung-cancer pathology. Combinatorial complexity quickly prohibits an exhaustive search of all k-element subsets of transcript sequence responses, for even small values of k. We show, however, that just by moving from k=1 to k=2, and considering pairs of expression vectors in concert, we achieve interesting results and identify seemingly biologically important genes that were not identified by previous analyses. When the approach is bootstrapped to k=3 and then k=4, we find, for the dataset of [1], nine 4-tuples of transcript sequences that each yield 97% correct classification for the five-class problem. We show that the same method applied to the dataset of [5] yields five 4-tuples of transcript sequences that each yield at least 97% correct classification among the six classes of lung cancer tumor contained in that dataset. Again, these results can be seen as

validating the method of Li et al. [8] presented at CAMDA 2000; [8] use a genetic algorithm to heuristically explore the space of  $k$ -element subsets. For the data dimensionality and the small value of  $k$  considered here, we were able to find the best subsets exactly, using exhaustive search. However, for larger datasets or larger  $k$ , a genetic algorithm or other heuristic search approach such as [8] use, becomes appropriate.

### 1.1 k-Nearest Neighbor Classifiers

The  $k$ -nearest-neighbor classifier, first introduced by Fix and Hodges [3], is the simplest and best-known non-parametric classifier. It is based on a distance, or dissimilarity measure  $d$ , that is assigned to all pairs of observations; for this study we used the L1 metric, where if  $x=(x_1, \dots, x_n)$ , and  $y=(y_1, \dots, y_n)$  are observations, then  $d(x,y)=\sum |x_i - y_i|$ .

The  $knn$  classifier is typically defined as follows. Suppose training data  $T= \{ t_1, \dots, t_r \}$  are a set of observations labeled by their class labels from  $C= \{ c_1, \dots, c_m \}$ . Let  $x$  be an observation whose class label is unknown. Define  $S_x \subseteq T$  to be the set of  $x$ 's  $k$  closest neighbors according to the distance metric  $d$  in  $T$ . Assuming no ties, let  $c_i$  be the class label that appears most frequently in the set  $S_x$ . Then  $x$  is assigned the class label  $c_i$  (notice in the case that there are 2 classes and  $k$  is odd, there will be no ties). In the case that more than one class label appears with equal frequency in  $S_x$ , we regress to the  $k-1nn$  classifier; if there is again a tie we regress to the  $k-2nn$  classifier, and so on. This must result in a unique class name, no matter how many classes there are, since when  $k=1$  there can be no ties.

Given the raw 12,600-dimensional dataset provided by [1], patients were divided into 5 groups, with each patient assigned to the group corresponding to his index mod 5 (Since the patients were grouped by class in the data set this was not a random partition, but rather had the effect of spreading out the number of patients of each class in each group as close to evenly as possible). We first show that, without any re-normalizing, pre-processing, or scaling, the 5-nearest-neighbor classifier correctly classifies 94% of the patients by lung tissue type in a 5-fold leave one out cross validation (see Table 1). From this we conclude that the five class problem of separating tissue samples into adenocarcinomas, squamous, SCLC, pulmonary carcinoid, and normal lung, is at most a problem of moderate difficulty. This is not surprising, given that the different classes are considered clinically distinct [1].

**Table 1. KNN five-fold cross validation on the entire 12,600-dimensional data set.**

	1 kNN % correct	3 kNN % correct	5 kNN % correct	7 kNN % correct
<b>Group 1</b>	95.1219	92.6829	95.1219	92.6829
<b>Group 2</b>	85.3659	87.8049	85.3659	85.3659

<b>Group 3</b>	90.2439	90.2439	92.6829	90.2439
<b>Group 4</b>	90.0000	97.5000	97.5000	90.0000
<b>Group 5</b>	95.0000	97.5000	100.0000	92.5000
<b>Average</b>	91.1330	93.5961	94.0887	90.1478

## 2. TWO CLASS SUB-PROBLEMS

We can also ask about the best genes for distinguishing each of the five individual classes from their complements. These problems vary in difficulty. In particular, there are six different transcript sequences that individually separate the pulmonary carcinoids from the other classes with 100% accuracy, and another three that achieve 99.5% accuracy, using 1nn in a leave-one-out cross validation. Similarly, the probe set 41231\_f\_at – described as high-mobility group (non-histone chromosomal) protein 17, gives complete separation with a 1nn between the SCLC class and all the rest, while an additional five transcript sequences give > 99.5% correct classification. The best individual sequence to separate the normal samples from all tumorous classes achieves 99% classification using the same method, while the top ten sequences all achieve > 97% classification. For the squamous class, there is one individual transcript sequence that achieves 98% correct classification (it is probe set 31791\_at, tumor protein 64 kDa with strong homology to p53, previously known to be a signature for squamous tumors [1]), and then there is a gap in discriminatory power; however, the top ten individual transcript sequences all achieve > 93% classification. The difficulty of the dataset, therefore, comes from the adenocarcinoma class: the best individual transcript sequence still achieves slightly less than 81% correct classification when separating the adenocarcinomas from the other four classes using 5-nearest neighbor.

A list of the top individual probe sets and their corresponding gene names for classifying each of these tumor types is omitted from this extended abstract for space reasons, but can be found at <http://www.cs.tufts.edu/~cowen/camda>.

## 3. IDENTIFYING GENES THAT JOINTLY DISCRIMINATE

Constructing a good classifier is not the only problem one wants to solve with microarray data, particularly for this problem where the different classes are clinically distinct. A more interesting biological problem is to identify a small subset of genes whose expression signatures themselves distinguish among different classes. Then these genes can be hypothesized to be involved biologically in the cancer pathology of the cell. We discuss this problem now.

As discussed in Section 1, both prior studies on the multiple classes of lung cancer [1,5] began by pre-filtering individual transcript sequences to come up with a subset of relevant genes.

In contrast, Li et al [8], suggested that additional power in feature selection for microarray data can be obtained by considering small subsets of transcript sequences that jointly discriminate. This approach has provably more power; however the computational obstacles grow quickly as the size of these subsets grow. In particular, to examine all pairs of transcript sequences in the 12,600-dimensional array requires 79,373,700 significance calculations, while to examine all triples of transcript sequences requires 333,316,624,2000 significance calculations. For this reason, [8] suggest a genetic algorithms approach to intelligently search this space, and give results and sensitivity of their methods to initial starting conditions in [7].

We take a more elementary bootstrapping approach to identify these joint discrimination sets as follow. First we examine all unique pairs of transcript sequences in the dataset, and retain the 1024 best pairs. Then those 1024 best pairs are matched with all unique third transcript sequences in the dataset, and the best 512 triples are maintained. Finally, the strongest 512 triples are matched with all unique fourth transcript sequences to obtain the best 4-dimensional classifier.

In the above description, “best” must be determined based on some measure of discriminatory power. We call upon our old friend k-nearest-neighbor again, and look at the percentage of correct classification based on a 1-nearest-neighbor classifier in a leave-one-out cross validation to determine the quality of each low-dimensional projection. We first tried the method on the dataset of [1]. As we raised the dimensionality of the model, the classification rate improved. The best of the transcript sequence pairs was capable of classifying 89% (182/203) of the observations correctly in a leave-one-out cross validation. Of the three-dimensional classifiers examined, six were found capable of correctly classifying 94% of the observations. Finally, of the four-dimensional models examined, nine 4-tuples were found that were capable of correctly classifying 97% (197/203) or more of the observations. The list of the 12 most frequent probe sets occurring in the set of the 512 strongest triples appears in Table 2. The set of the nine best 4-tuples appears in Table 3. Information about the biological significance of some of these genes appears in Section 4. A longer version of Table 2 and more biological information can be found at <http://www.cs.tufts.edu/~cowen/camda>.

**Table 2 Frequently occurring transcript sequences among top triples, with their frequency in the top triples and pairs.**

Probe set	Frequency in top 512 triples	Frequency in top 1024 pairs
1814	273	197
41325	108	161
31791	59	10
36160_s	48	18
36148	37	7

37398	27	23
38174	26	24
37182	20	4
33904	19	13
36133	16	19
38032	16	12
35868	15	28

(We remark that two of the best 4-tuples do very poorly on the SCLC class; whereas we showed that distinguishing SCLC observations from the others was among the easiest of the two-class problems. Thus, by separately classifying non-SCLC and SCLC observations first, we could improve the classification rate to 98.5% (200/203). Alternately, combining one of these best 4-tuples (the second one in Table 2) with the single probe set 39990\_at (a good classifier for SCLC), results in a 5-dimensional subset for which 1-nearest neighbor classifies 98% (199/203) correctly.)

**Table 3 List of the nine best 4-dimensional transcript sequence classifiers.**

classifier	AD (139)	NL (17)	SC LC (6)	SQ (21)	COID (20)	Total
3814, 1814, 33529, 1071	138	17	3	20	20	97.5%
37302, 41325, 31791, 763	136	16	6	19	20	97%
31791, 41325, 36174, 40223r	137	17	3	20	20	97%
31791, 41325, 35595, 33218	137	16	6	18	20	97%
36148, 37391, 33218, 37991	137	16	4	20	20	97%
37302, 41325, 31791, 41245	137	16	6	18	20	97%
1814, 185, 36139, 39990	137	16	6	18	20	97%
37302, 41325, 31791, 32240	136	16	6	19	20	97%
31791, 41325, 39158, 38004	136	16	6	19	20	97%

The nine best 4-tuples in Table 3 contain within them 22 unique transcript sequences; of these 31791\_at, 41325\_at, 33218\_at, 37302\_at, and 1814\_at occur multiple times across the nine 4-tuples (in fact, 31791\_at and 41325\_at occur as a pair in six of the

nine best 4-tuples). This argues for the biological importance of these genes in lung-cancer, particularly those that occur multiple times on the list, and in fact 31791\_at is tumor protein 63 kDa with strong homology to p53, involved in cell growth regulation, known to be involved in lung cancer pathology, and previously identified as critical by both [1] and [5]. On the other hand, based on these results, we conjecture that 41325\_at, identified as potassium channel, subfamily K, member 3, that encodes one of the superfamily of potassium channel proteins [2] is also of biological importance, and this gene was not flagged in any previous study. The biological meanings of the other probe sets that make up the classifiers in Table 3 is discussed briefly in Section 4, more details are omitted for space, but can be found at <http://www.cs.tufts.edu/~cowen/camda>.

To validate the method, we turned to the second data set of Garber et al. [5] We show that our method has similar performance on the six-class problem of dataset 2. Table 4 shows the performance of the top five 4-dimensional classifiers by gene accession number, each of which correctly classifies 58 of the 59 patients, greater than 98% correct classification rate on the five-class problem considered.

Once again, we postulate that the genes that show up multiple times in this table have biological significance for lung cancer pathology. In particular, we discuss what is known about R70462, H65075 and T84152 in Section 4.

**Table 4. List of the five best 4-dimensional transcript sequence classifiers.**

classifier	AD (34)	LC LC (4)	NL (5)	SC C (12)	SC LC (4)	Total
R70462, H97677 R26186, AA007308	34	3	5	12	4	98.3%
R70462, AA862435, H65065, T84152	34	4	5	11	4	98.3%
R70462, T47454, N55459, AA460571	34	3	5	12	4	98.3%
R70462, H02848, H65065, H77706	34	3	5	12	4	98.3%
R70462, AA186348, H6505, T84152	33	4	5	12	4	98.3%

#### 4. BIOLOGICALLY SIGNIFICANT GENES

We postulate that the transcript sequences that occur most frequently in tables 2, 3, and 4, are biologically significant in lung cancer pathology. A full description of what is known about these genes for both datasets appears in supplementary information at

<http://www.cs.tufts.edu/~cowen/camda> -- for space reasons, we only give some highlights in this extended abstract.

Two of the probe sets we find are also explicitly identified not only by our methods, but also in the papers of [1] and [5]. These are probe set 1814 (transforming growth factor, beta receptor II), and probe set 31791 (tumor protein 63 kDa with strong homology to p53). Both are known to be involved in the pathology of multiple cancers [6,11].

We additionally find the probe set 33218, which occurs in one of the top 4-tuples for the dataset of [1], is the same as R70462, which occurs in all the top triples in the dataset of [5]. The paper of [1] does not list this as an important gene at all; in the paper of [5] it is on a list of nearly 500 genes that they identify as having a high expression value in all adenocarcinomas, but a low expression value in all squamous samples. The gene is v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian). It encodes a tumor antigen, p185, which is serologically related to EGFR, the epidermal growth factor receptor [14]. Its role in cancer has been studied in several other papers, for example, [13] found that its over-expression in cancers corresponds to poor prognosis, enhanced metastatic potential, and chemoresistance.

Most of the frequently occurring transcript sequences in tables 2, 3, and 4 are not identified as important in either of the papers of [1] and [5]. However, T84152, caveolin 2, has recently been implicated to have some role in cancer in the biology literature. A 2003 paper of Fong et al. [4] shows a positive correlation of the expression of caveolin 1 and caveolin 2 with tumor grade and squamous features of urothelial carcinoma. They suggest that caveolin 1 and caveolin 2 be studied further to determine a possible role in tumor progression and squamous differentiation. Other genes that appear important in our analysis but have not been previously identified as such by [1] and [5] are 41325\_at, identified as potassium channel, subfamily K, member 3, that encodes one of the superfamily of potassium channel proteins [2] and H65065, visinin-like 1, also referred to as VILIP1, which Lin et al. [9] show modulates the surface expression and agonist sensitivity of the alpha 4 beta 2 nicotinic acetylcholine receptor in response to changes in levels of calcium. Minna [10] links the alpha 4 beta 2 acetylcholinic receptors to lung cancer directly, claiming that smoking addiction is a result of the action of nicotine on these receptors.

#### 5. CONCLUSIONS

We have shown that the simplest non-parametric classifiers can have some utility for some microarray classification problems, acting on the entire non-dimension reduced dataset. For the problem of determining small sets of transcript sequences that have discriminatory power (and thus possible significance in the biological pathway), we show that increasing the dimensionality of these sets (considering pairs, triples or four-tuples, rather than individual transcript sequences one by one) can lead to significant

improvements with each dimension gained. As a result, we caution the practitioner against reducing the dimensionality of the data too quickly.

## 6. ACKNOWLEDGMENTS

Thanks to Donna Slonim for her encouragement and for reading an early draft of this paper.

## 7. References

- [1] A. Bhattacharjee, W. Richards, J. Shaunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Fillette, M. Loda, G. Weber, E. Mark, E. Lander, W. Wong, B. Johnson, T. Golub, D. Sugarbaker, and M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98, 24, 13790-13795, November 2001.
- [2] F. Duprat, F. Lesage, M. Fink, R. Reyes, C. Heurtenaux, and M. Lazdunski, TASK, a human background K<sup>+</sup> channel to sense external pH variations near physiological pH, *Ebo J.*, 16, 5464-6471, 1997.
- [3] E. Fix and J. Hodges, Discriminatory analysis: Nonparametric discrimination: Consistency properties. Technical report 21-49-004, USAF School of Aviation Medicine, 1951.
- [4] A. Fong, E. Garcia, L. Gwynn, M. Lisanti, M. Fazzari, and N. Li, Expression of Caveolin-1 and Caveolin-2 in urothelial carcinoma of the urinary bladder correlates with tumor grade and squamous differentiation, *Am J Clin Pathol* 120(1):93-100, 2003.
- [5] M. Garber, O. Troyanskaya, K. Schluens, S. Petersen, Z. Thaessler, M. Gengelbach, M. van de Rijn, G. Rosen, C. Perou, R. Whyte, R. Altman, P. Brown, D. Botstein, and L. Petersen, Diversity of gene expression in adenocarcinoma of the lung, *PNAS*, 98(24): 13784-13798, November 2001.
- [6] K. Hibli, B. Trink, M. Patturajan, W. Westra, O. Caballero, D. Hill, E. Ratovitski, J. Jen and D. Sidransky, AIS is an oncogene amplified in squamous cell carcinoma, *PNAS*, 97: 5462-5467, 2000.
- [7] L. Li, C. Wienberg, T. Darden, and L. Pedersen. Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method, *Bioinformatics*, 17(12): 1131-1142, 2001.
- [8] L. Li, P. Bushel, L. Pedersen, T. Darden, H. Hamadeh, L. Bennett, C. Afshari, R. Paules, D. Umbach, and C. Weinberg, Computational analysis of Leukemia microarray expression data using the GA/KNN method and other existing tools, in *Methods of Microarray Data Analysis: Papers from CAMDA 2000*, S. Lin and K. Johnson, eds., Boston: Kluwer Academic Publishers, 2001.
- [9] L. Lin, E. Jeanclos, M. Treuil, K. Braunewell, E. Gundelfinger, and R. The calcium sensor protein visinin-like protein-1 modulates the surface expression and agonist sensitivity of the alpha 4beta 2 nicotinic acetylcholine receptor. *J Biol Chem* 1;277(44): 41872-8, 2002.
- [10] J. D. Minna, Nicotine exposure and bronchial epithelial cell nicotinic acetylcholine receptor expression in the pathogenesis of lung cancer. *J Clin Invest.* 111(1): 31-33, 2003.
- [11] S. Markowitz, J. Wang, L. Myeroff, R. Parson, L. Sun, J. Lutterbaugh, R. Fan, E. Zborowska, K. Kinzler, B. Vogelstein, M. Brattain, and J. Wilson, Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability, *Science*, 268: 1336-1338, 1995.
- [12] D. Slonim, From patterns to pathways: gene expression data analysis comes of age, *Nature Genetics*, 32(S), 502-508, 2002.
- [13] M. van de Vijver, J. Petersen, W. Mooi, P. Wisman, J. Lomans, O. Dalesio and R. Nusse, NEU-protein overexpression in breast-cancer: association with comedo-type ductal carcinoma in situ and limited prognostic value in stage II breast cancer, *New England Journal of Medicine*, 319: 1239-1245, 1988.
- [14] T. Yang-Feng, A. Schechter, R. Weinberg, U. Francke, Oncogene from rat neuro/glioblastomas (human gene symbol NGL) is located on the proximal long arm of human chromosome 17 and EGFR is confirmed at 7p13-q11.2 *Cytoget. Cell Genetics* 40: 784, 1985.