



Higher Dimensional Approach for Classification of Lung Cancer Microarray Data

Nathan Palmer

Tufts University / MIT

**(Joint work with Frederick Crimins, Robert
Dimitri, Tsvika Klein and Lenore J. Cowen)**

Outline

- Classification of Tissue Types
- Gene Selection for Class Prediction
- Biological Significance of Reported Genes

Dataset: 203 Tissue Samples

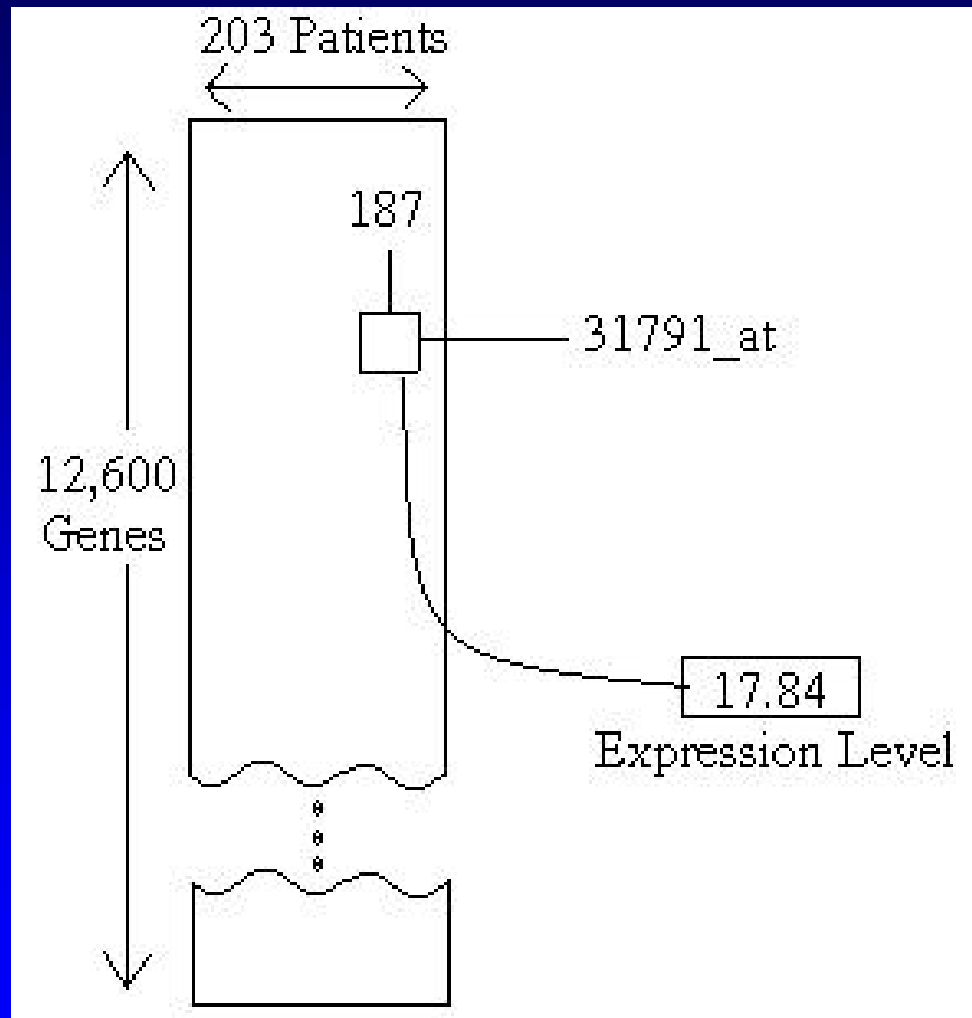
Expression values for 12,600 transcript sequences, or genes, for each of:

- 186 cancer tissue samples classified as:
 - Adenocarcinomas (139)
 - Squamous cell lung carcinomas (21)
 - Pulmonary carcinoids (20)
 - Small-cell lung carcinomas (SCLC) (6)
- 17 normal tissue samples

Bhattacharjee et al (2001) PNAS

Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses

Dataset: 203 Tissue Samples



Outline

Classification of Tissue Types

Selecting a Classifier

Interpreting the Data

Classification of Tissue Types

Problem

Given: Tissue samples with expression data, labeled by cancer type (or normal). This is called a *training set*.

Determine: Rule to assign cancer type to a new, unlabeled tissue sample based on its expression data.

Two Classification Problems

The 5-Class Problem:

Allow known tissue samples to be classified as any one of 4 cancer types, or normal tissue.

Try to place a new, unlabeled tissue sample into one of these 5 classes

Two Classification Problems

The 2-Class Problem:

Consider only 1 type of cancer (or normal) tissue; Allow known tissue samples to be classified as either members of this class, or not.

Try to determine whether or not a new, unlabeled tissue sample is of this type.

Example:

Determine whether or not a new tissue sample is a SCLC.

Selecting a Classification Rule

k -Nearest Neighbor Classifiers:

- Fix k as a constant.
- Given a new tissue sample, x , use a dissimilarity (distance) metric to select the k tissue samples in the training set that are “closest” to x .
- Assign to x the tissue type most frequently appearing in those k nearest tissue samples.

Selecting a Classification Rule

Defining a Distance Metric:

Each tissue sample is associated with 12,600 real-valued expression levels.

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{12600} \end{pmatrix} \quad a_i \hat{I} \hat{A}$$

Selecting a Classification Rule

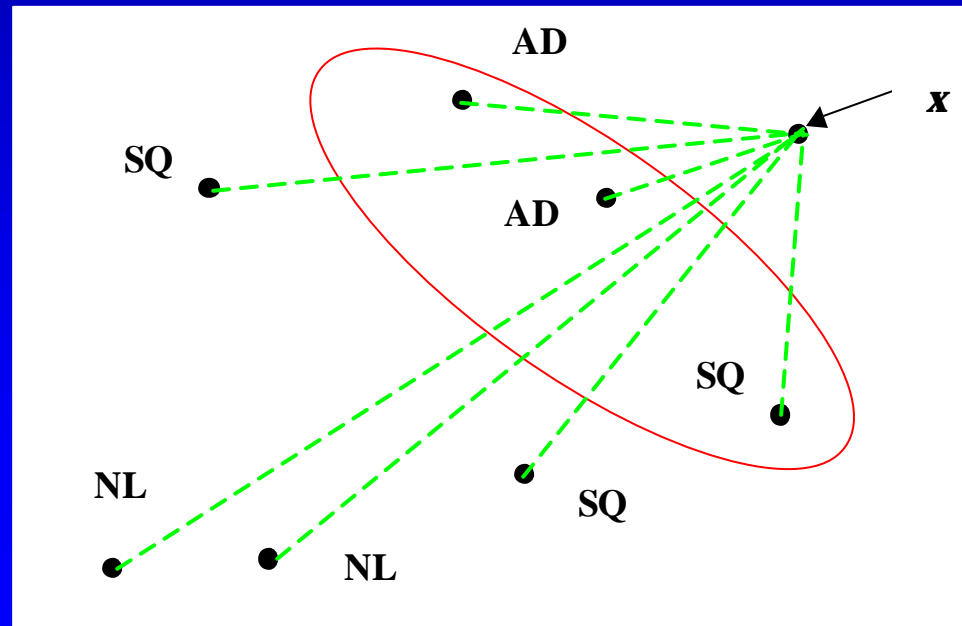
Defining a Distance Metric:

Treat each tissue sample as a 12,600-dimensional real-valued vector and use Euclidean distance as our distance metric.

Selecting a Classification Rule

k -NN example, considering only 2 genes, $k = 3$:

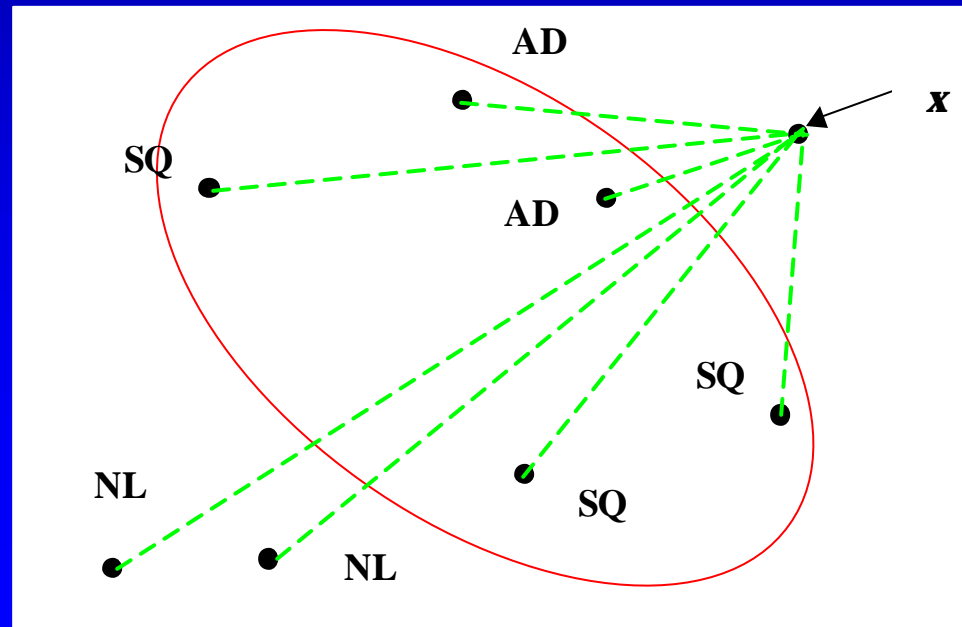
x gets classified
as
adenocarcinoma



Selecting a Classification Rule

k -NN example, considering only 2 genes, $k = 5$:

x gets classified
as squamous

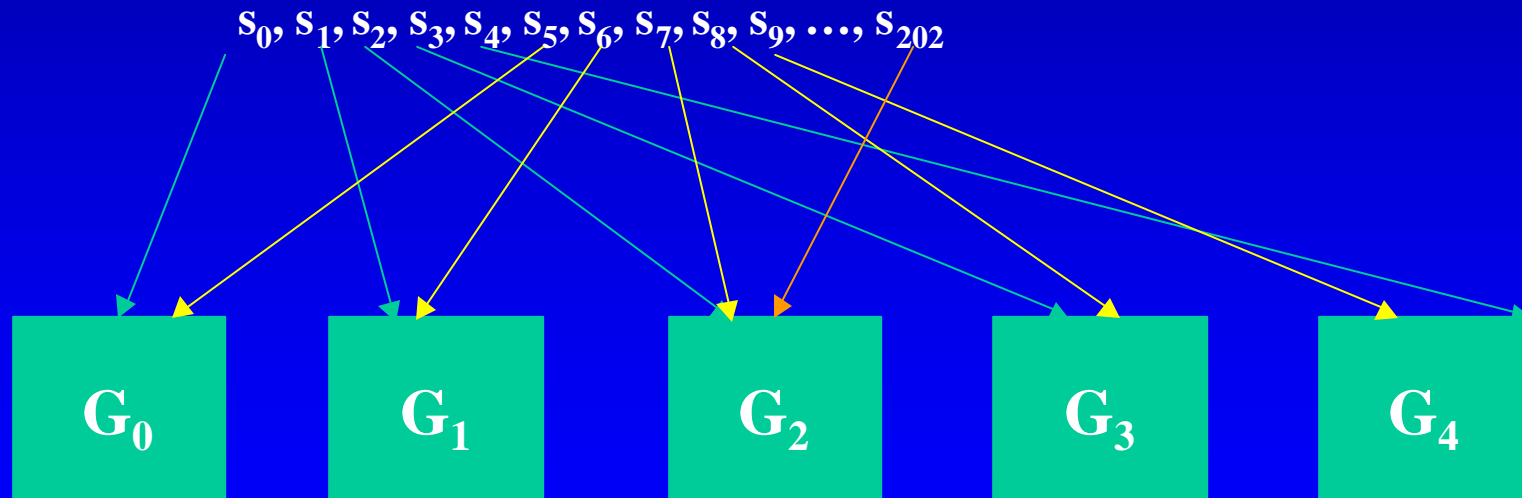


Can k-NN Separate These Tissue

Types?

An initial experiment:

For the purpose of cross-validation, divide the 203 tissue samples into 5 groups. Assign each sample to group G_i , where $i = \text{sample index mod } 5$.



Five-Fold Cross-Validation

For $k = \{1,3,5,7\}$

classify this group

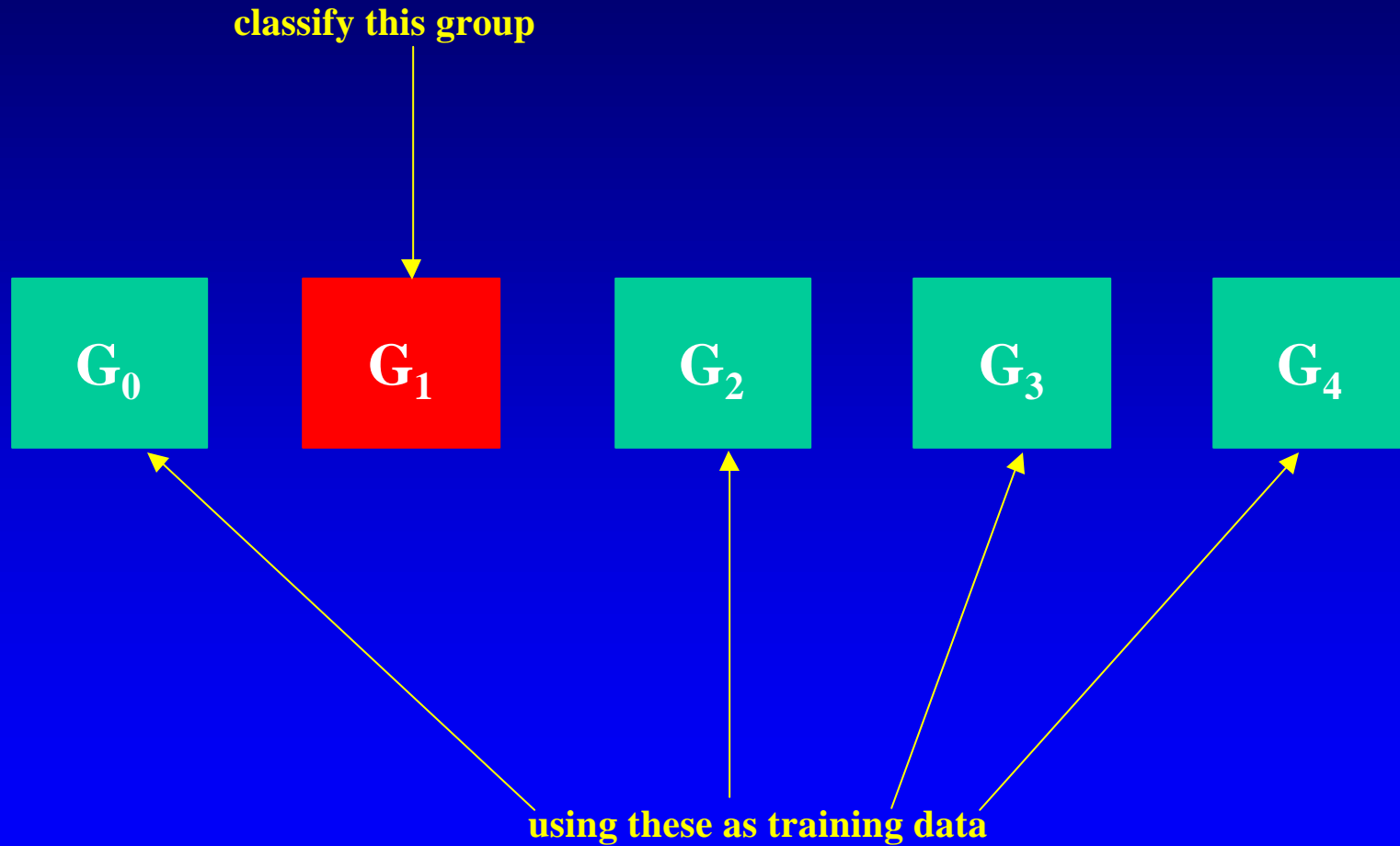


using these as training data



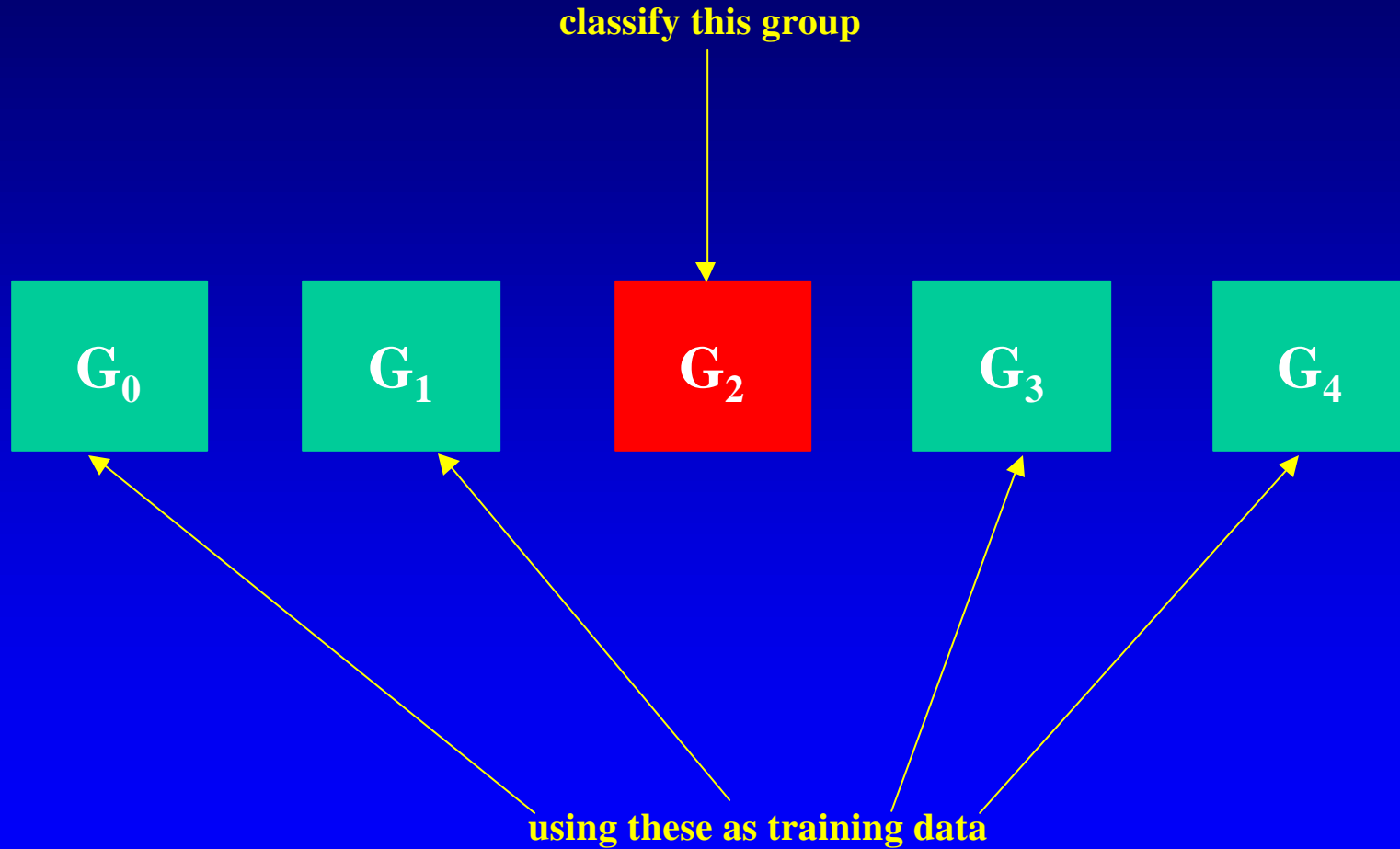
Five-Fold Cross-Validation

For $k = \{1,3,5,7\}$



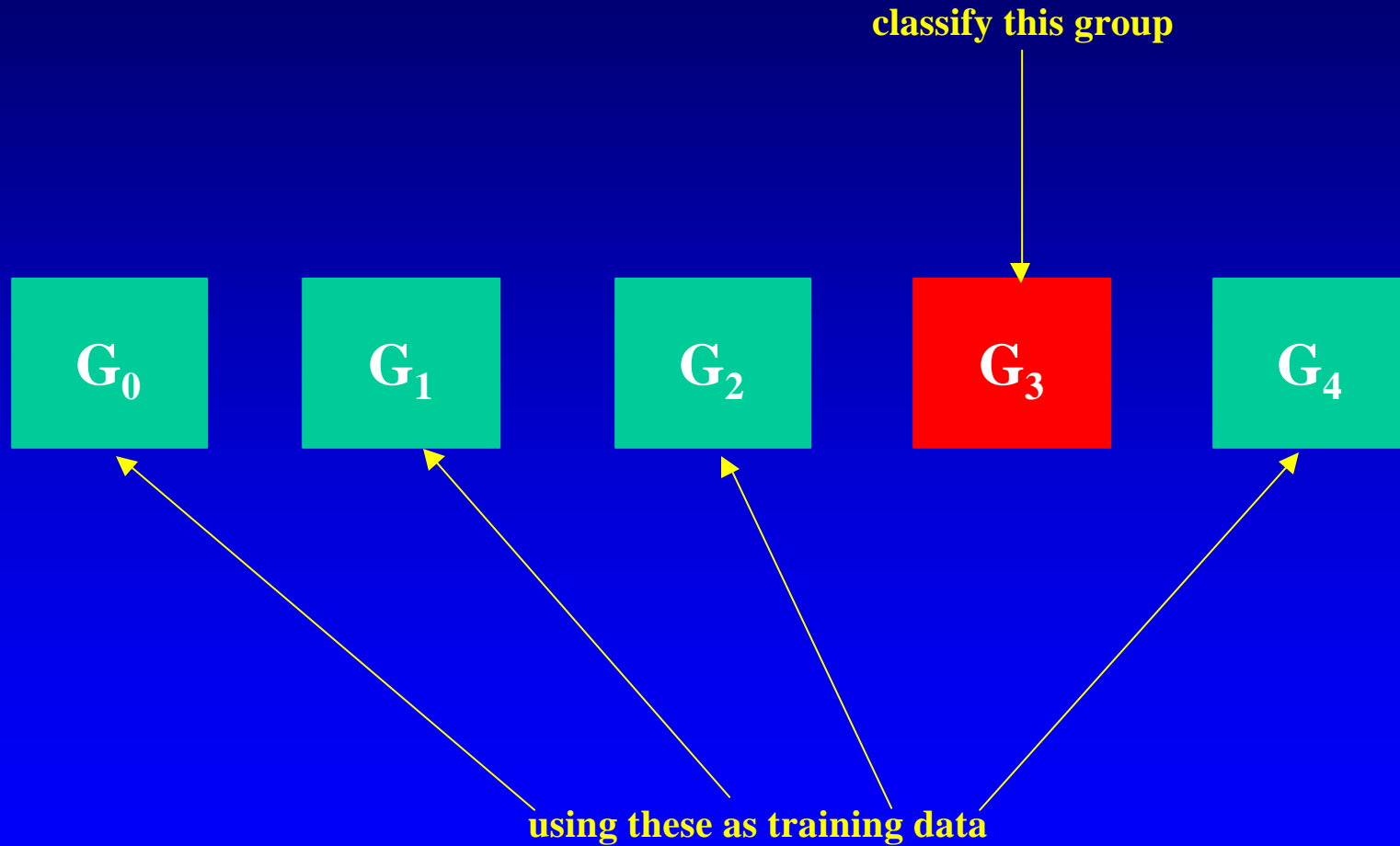
Five-Fold Cross-Validation

For $k = \{1,3,5,7\}$



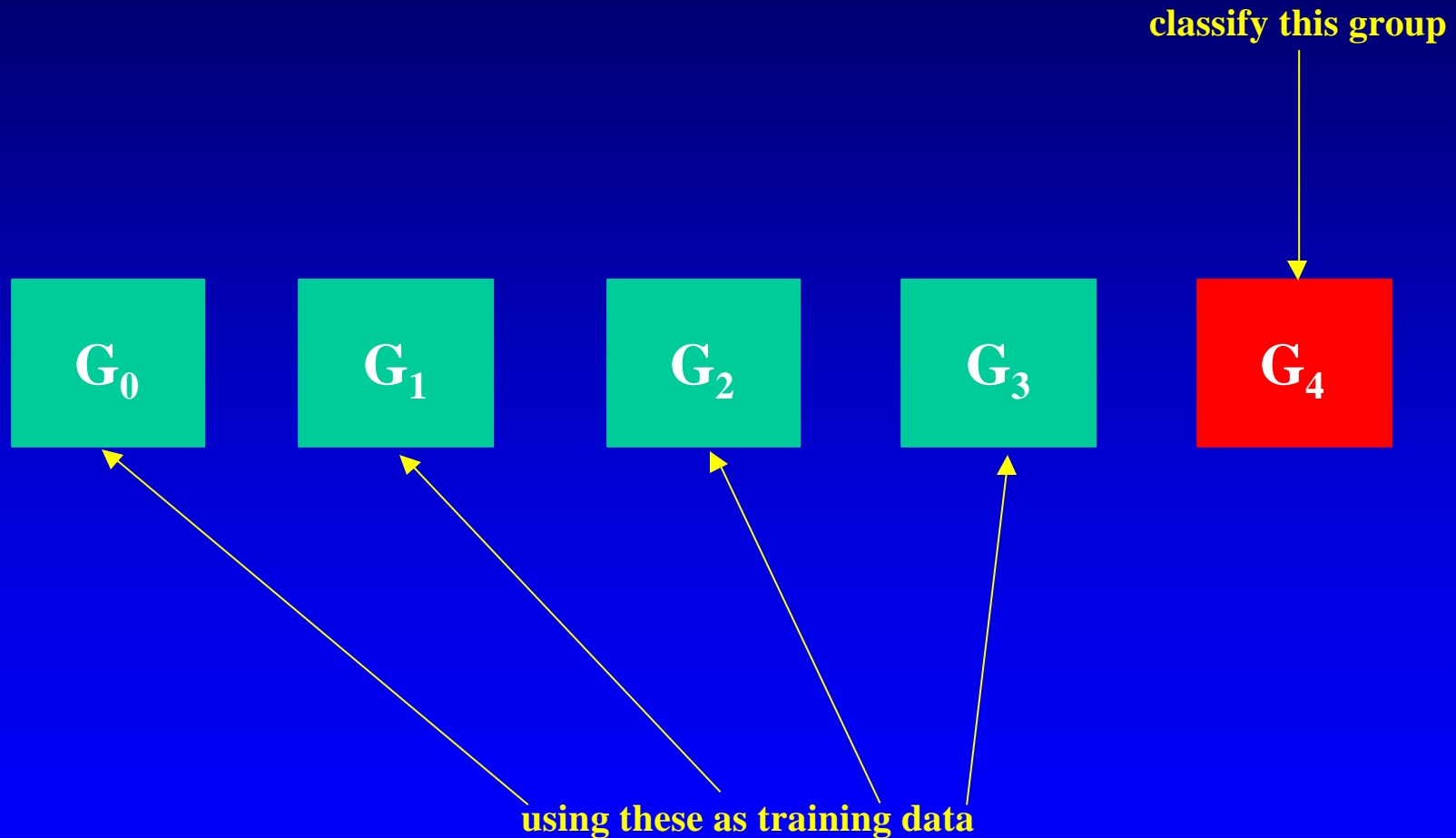
Five-Fold Cross-Validation

For $k = \{1,3,5,7\}$



Five-Fold Cross-Validation

For $k = \{1,3,5,7\}$



Five-Fold Cross Validation Results

	1NN % correct	3NN % correct	5NN % correct	7NN % correct
Group 1	95.1219	92.6829	95.1219	92.6829
Group 2	85.3659	87.8049	85.3659	85.3659
Group 3	90.2439	90.2439	92.6829	90.2439
Group 4	90.0000	97.5000	97.5000	90.0000
Group 5	95.0000	97.5000	100.0000	92.5000
Average	91.1330	93.5961	94.0887	90.1478

*k*NN five-fold cross validation on the entire 12,600-dimensional data set of Bhattacharjee et al

Results

Conclusion:

The problem of differentiating between adenocarcinoma, squamous, SCLC, pulmonary carcinoid, and normal lung tissue samples is not that hard!

Outline

Gene Selection for Class Prediction

- Identifying Marker Genes for Each Tissue Type
 - Identifying Genes that Jointly Discriminate

Identifying Marker Genes for Each Tissue Type

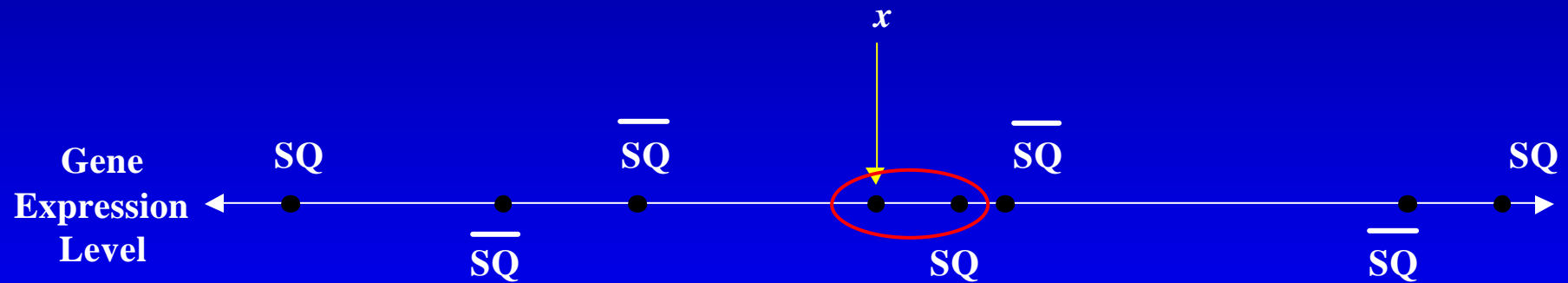
Goal: Find genes that separate each tissue type from the rest of the dataset.

Identifying Marker Genes for Each Tissue Type

Approach: Evaluate each gene using 1NN in a leave-one-out cross-validation.

Identifying Marker Genes for Each Tissue Type

Example: using 1NN to evaluate a gene's ability to separate the squamous class



***x* gets labeled as a squamous tissue, since its nearest neighbor, by this gene, is a squamous tissue**

Identifying Marker Genes for Each Tissue Type

Pulmonary Carcinoid:

6 genes separate with 100% accuracy

Identifying Marker Genes for Each Tissue Type

SCLC:

Gene 41231_f_at (high-mobility group (non-histone chromosomal) protein 17) separates with 100% accuracy.

5 other genes separate with 99.5% accuracy.

Identifying Marker Genes for Each Tissue Type

Squamous:

Gene 31791_at (tumor protein 64 kDa with strong homology to p53, previously known to be a signature for squamous tumors*) separates with 98% accuracy.

**Bhattacharjee et al (2001) PNAS*

Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses

Identifying Marker Genes for Each Tissue Type

Adenocarcinoma:

9 genes separate with better than 77% accuracy.

Taking the best gene and using 5NN, we still get slightly less than 81% accuracy.

Identifying Marker Genes for Each Tissue Type

Conclusion:

The adenocarcinomas present the greatest challenge in this dataset.

Outline

Gene Selection for Class Prediction

- ✓ Identifying Marker Genes for Each Tissue Type
 - Identifying Genes that Jointly Discriminate

Identifying Genes that Jointly Discriminate

Goal: Find small subsets of genes that distinguish between the tissue types.

Identifying Genes that Jointly Discriminate

- Motivation:
- Improve classification by reducing noise.
 - Uncover possible biological interactions between genes.

Identifying Genes that Jointly Discriminate

Computational obstacles grow exponentially as we increase the size of the subsets we examine.

For example, $\binom{12,600}{2} = 79,373,700$

$$\binom{12,600}{3} = 333,316,624,200$$

Identifying Genes that Jointly Discriminate

Method:

Examine all unique pairs of genes in the dataset, retaining the 1024 best pairs.

Match those 1024 pairs with all unique third genes, retaining the best 512 triplets.

Finally, match those 512 triples against all unique fourth genes to obtain the best 4-dimensional classifiers.

Identifying Genes that Jointly Discriminate

Examine the percentage of correct classifications based on 1NN in a leave-one-out cross validation.

Identifying Genes that Jointly Discriminate

Results:

1 pair capable of 89% correct classification,

3 triplets capable of 94% ,

9 quartets capable of ≥ 97 %

List of 9 Best 4-Dimensional Gene Classifiers

classifier (probe set)	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%

Frequently Occurring Genes

Gene (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

Method Validation: Garber Dataset

classifier (accession)	AD (34)	LC LC (4)	NL (5)	SCC (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%

List of the five best 4-dimensional transcript sequence classifiers (by gene accession number) from the data set of Garber et al.

Outline

Biological Significance of Reported Genes

Biologically Significant Genes

Previously Identified Genes:

- probe set 1814 (transforming growth factor, beta receptor II)
- probe set 31791 (tumor protein 63 kDa with strong homology to p53)

Both are identified by Bhattacharjee and Garber, known to be involved in the pathology of multiple cancers.*+

* 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.

+ 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.

Biologically Significant Genes

Previously Identified Genes:

- probe set 33218 = R70462

Garber notes high expression level in adenocarcinomas, low in squamous, Bhattacharjee does not identify this gene. Over-expression in cancers has been previously linked to poor prognosis and chemoresistance.*

* 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.

Biologically Significant Genes

Previously Unidentified Genes:

- T84152 (caveolin 2)

Fong et al. show a positive correlation of the expression of caveolin 1 and caveolin 2 with tumor grade and squamous features of urothelial carcinoma.*

* 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.

Biologically Significant Genes

Previously Unidentified Genes:

- 41325_at (potassium channel, subfamily K, member 3)

Encodes one of the superfamily of potassium channel proteins which Lin et al.* show modulates the surface expression and agonist sensitivity of the alpha 4 beta 2 nicotonic acetylcholine receptor. Minna⁺ 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.

* 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.

+ 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.

Conclusion

- We present a simple, tractable algorithm for selecting small subsets of genes that jointly discriminate the tissue types with high accuracy.
- Preprocessing / filtering is not necessary to uncover signal in this data.