Unsupervised Clustering of Colorectal Cancer Gene Microarray Data

Will Cukierski[1,2], Wenjin Chen[1], Huiqi Chu[1], David J. Foran[1,2]

1. University of Medicine and Dentistry of New Jersey, Cancer Institute of New Jersey. 2. Rutgers University

Abstract

Gene microarrays have been studied as a means to find relevant genes within a large set of candidates. This selection is typically accomplished through clustering algorithms that utilize a variety of statistical similarity measures. As opposed to gene selection, this study explores a colorectal cancer gene microarray to find clinically-meaningful groups of patients (as judged from pathology reports). The existence of a pre-cancerous class is investigated. Several methods of dimensionality reduction, classification, and unsupervised clustering are performed and compared within the framework of gene microarrays.

Data

- 22,283 mRNA features from an Affymetrix oligonucleotide array (details are given in [1, 4])
- 360 samples of six tissue types
- Pathology reports: T1CC = T1 staged colon cancer ... T4CC = T4 staged colon cancer, CRY = Crohn's test positive sample, CRN = Crohn's test negative sample

Tissue Type	# Samples	Mean Correlation
Colon Cancer (CC)	174	0.92
Primary Polyp (PP)	50	0.94
Liver Metastasis (LIM)	47	0.91
Lung Metastasis (LUM)	20	0.93
Normal Mucosa (NM)	49	0.92
Normal Liver (NLI)	13	0.94
Normal Lung (NLU)	7	0.96

Table 1: Correlation coefficients are the intra-class Pearson coefficients, $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Dimensionality Reduction

- Nonlinear dimensionality reduction methods are better suited for microarray analysis [3].
- Principle component analysis (PCA), a linear technique which seeks to preserve the variance in the high-dimensional data, was compared with the nonlinear isometric feature mapping algorithm (ISOMAP) of Tenenbaum et al. [5].
- ISOMAP retains a global ordering which is lost in the PCA embedding.

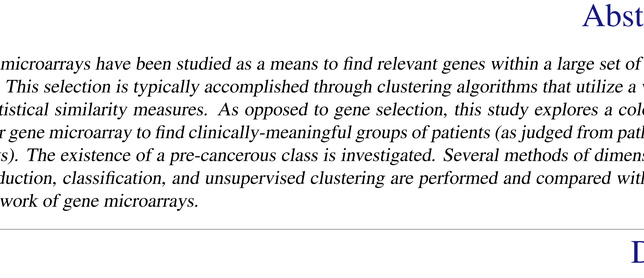

Figure 1: 2D embedding of the colon cancer, polyp, and normal groups. ISOMAP shows the transition from normal to polyp to cancer. PCA fails to separate the classes. Labels are the ground truth from the pathology reports.

- Due to the curse of dimensionality, Euclidean distance has limited meaning in the 22,283-dimensional space of RNA measurements.

Acknowledgements

This research was funded, in part, by a grant from the NIH through contract 263-MQ-610681 from the National Cancer Institute.

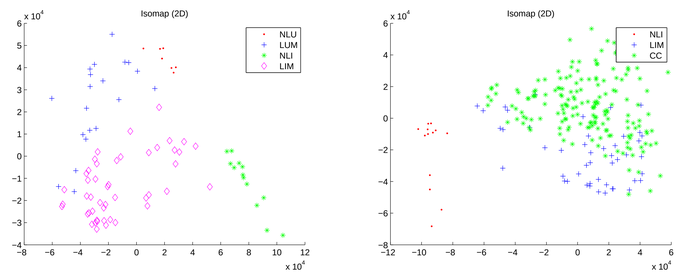

Figure 2: The primary cell of origin is visible in the embedded dataset. Biopsies taken from liver and lung metastases are embedded with samples of the primary colorectal tumor. The embedding of the liver metastases, in particular, shows a genetic profile which is colorectal in origin, but partially reflects the host environment (the samples are embedded between the cancer points and the normal liver points). It is unclear whether this effect is produced by contamination of the biopsy with small amounts of normal tissue, or genetic adaptation of the metastasis to the host environment.

Classification

Supervised classification was performed to assess the likelihood for success using unsupervised methods. If a trained classifier fails, it is unlikely unsupervised clustering will succeed. A polynomial Support Vector Machine (SVM)[6] classifier was trained using 1/3 holdout cross validation (repeated 100 times) on several partitions of the dataset. The classifier was unable to separate clinical stage or Crohn's reaction.

1	2	3	4	Dim.	Err(%)	Var(%)
CC	NM	-	-	22283	1.37	0.01
CC	NM	PP	-	22283	9.74	0.07
CC	NM	PP	LIM	22283	10.96	0.07
NLI	LIM	-	-	22283	2.32	0.14
NLU	LUM	-	-	22283	0.13	0.02
T1CC	T2CC	T3CC	T4CC	22283	45.98	0.30
T2CC	T4CC	-	-	22283	35.06	1.18
CC	LIM	LUM	-	22283	7.80	0.12
CRY	CRN	-	-	22283	31.53	0.26

Table 2: Accuracy of a polynomial SVM classifier for multi-class distinction.

k-Means Clustering

k-Means was run in the original 22,283 dimensional space, using correlation as the distance metric, for several partition numbers and permutations of sample types. k-Means provided mixed results in this dataset; clusters correspond to tissue type in cases where classes are separated and non-overlapping, but provides poor results for cases where the data overlap, where the clusters are not well separated, or when the number of samples per cluster is markedly different.

Hierarchical Clustering

A hierarchical classification method was implemented, using correlation as the distance metric. A hierarchical tree was constructed using the weighted pair group method using arithmetic averaging (WPGMA). The method initializes each sample as its own cluster, and merges the clusters with the highest correlation.

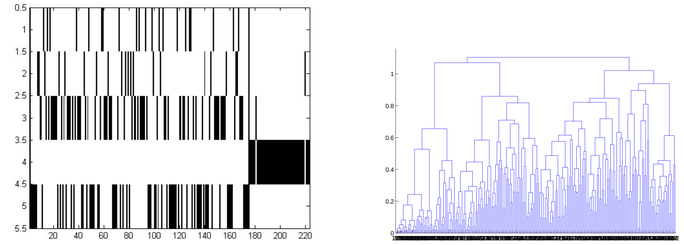

Figure 3: Hierarchical clustering results for the CC and NM groups. The vertical axis corresponds to the cluster number, the horizontal axis to each sample. A black bar indicates the cluster in which each sample is placed. Vertical black lines partition the samples into the ground-truth classes. The large black area to the right indicates a cluster of normal samples, while the cancer group is split across four other clusters. The figure on the right shows the dendrogram for this clustering.

Quantum Clustering

A novel method of clustering by Horn et al.[2], called Quantum Clustering, was also implemented. This physics-inspired method creates a probabilistic wave function and potential function, which constitute a solution to the Schrödinger equation. Singular value decomposition is first performed, then each data point is assigned to a Gaussian of width σ by a Parzen-window approach. One free parameter is varied and indirectly determines the number of clusters.

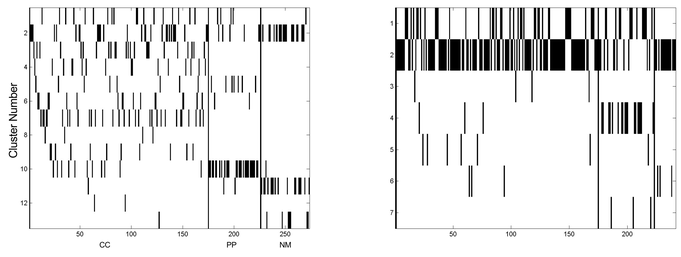

Figure 4: Left: Clustering results for the samples from 3 classes: CC, PP, NM. The PP group is clustered tightly into one group, the NM is split across 2 clusters, while the CC groups is distributed among many clusters. Right: Clustering results for the samples from CC, LIM and LUM.

Conclusions

1. Clinical stage does not appear to be reflected in the genetic profile.
2. Nonlinear distance metrics are necessary to capture the manifold shape on which the data resides.
3. Tumor cell of origin is visible in the expression profile.
4. Ongoing & future work: mine the clusters for samples with similar pathology, multimodal analysis using microarray data and radiological images

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proc Natl Acad Sci U S A, 96 (1999), pp. 6745–50. Journal Article United states.

[2] D. Horn and I. Axel, *Novel clustering algorithm for microarray expression data in a truncated svd space*, Bioinformatics, 19 (2003), pp. 1110–5. Journal Article England.

[3] G. Lee, C. Rodriguez, and A. Madabhushi, *An empirical comparison of dimensionality reduction methods for classifying gene and protein expression datasets.*, ISBRA 2007, LNBI 4463 (2007), pp. 170–181.

[4] D. A. Notterman, U. Alon, A. J. Sierk, and A. J. Levine, *Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays*, Cancer Res, 61 (2001), pp. 3124–30. Journal Article Research Support, Non-U.S. Gov't United States.

[5] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 290 (2000), pp. 2319–23. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United states.

[6] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, (1995).