Expression analysis of Barrett's epithelium and normal gastrointestinal tissues

Ka Yee Yeung
Michael T. Barrett
Jeff Delrow
Patricia L. Blount
Li Hsu
Walter L. Ruzzo
Brian J. Reid
Peter S. Rabinovitch

Technical Report UW-CSE-2000-11-01
November, 2000

Department of Computer Science & Engineering
University of Washington
Seattle, WA 98195

Expression analysis of Barrett's epithelium and normal gastrointestinal tissues

Ka Yee Yeung
Dept of Computer Science and Engineering, University of Washington
kayee@cs.washington.edu

Michael T. Barrett*
Fred Hutchinson Cancer Research Center
mbarrett@fhcrc.org

Jeff Delrow, Patricia L. Blount, Li Hsu
Fred Hutchinson Cancer Research Center

Walter L. Ruzzo
Dept of Computer Science and Engineering, University of Washington

Brian J. Reid
Fred Hutchinson Cancer Research Center

Peter S. Rabinovitch
Dept of Pathology, University of Washington

December 7, 2000

Abstract

Barrett's esophagus is a premalignant condition that arises due to chronic acid reflux in which the normal squamous epithelium of the esophagus is replaced by a metaplastic columnar epithelium. A fundamental question is the distinction between neoplastic Barrett's epithelium and surrounding normal tissues of the upper gastrointestinal tract. For example, although it arises in the esophagus, the Barrett's epithelium more closely resembles the epithelium of the duodenum at the histologic level. Therefore, we compared the transcriptional profile of the Barrett's epithelium to those of normal upper gastrointestinal tissues, including gastric epithelium, squamous epithelium of the esophagus and duodenal epithelium. We found that the Barrett's epithelium has comparable similarities to the three normal gastrointestinal tissues at the expression level. In addition, we proposed a novel approach to filter out non-tissue specific genes. We searched for tissue specific patterns, and identified many tissue specific genes. Furthermore, we developed a novel algorithm to identify genes that drive the similarity (or dissimilarity) between different tissue samples.

*Equal contribution to this work.

1 Introduction and Motivation

Barrett's esophagus is a metaplasia that develops as a complication in 10-20% of patients with chronic gastroesophageal reflux disease and predisposes to the development of adenocarcinomas of the esophagus and the gastric cardia ([Hamilton *et al.*, 1988], [Phillips and Wong, 1991]). Since the mid 1970s, the incidence of Barrett's-associated adenocarcinoma has increased more rapidly than that of any other cancer in the United States [Blot *et al.*, 1991]. Unfortunately, most patients who develop an esophageal adenocarcinoma present when the cancer is advanced and incurable, and more than 90% will die of their disease [Silverberg *et al.*, 1990]. Patients with Barrett's esophagus typically have symptoms of gastroesophageal reflux, such as heartburn or indigestion and they frequently seek medical attention before they develop cancer. The Barrett's epithelium can be safely visualized and biopsied during upper gastrointestinal endoscopy. At the present time, total removal of Barrett's epithelium requires esophagectomy, a procedure with substantial morbidity and mortality. However, a systematic protocol of endoscopic biopsies can detect early curable cancers arising in Barrett's esophagus. Therefore, the standard of care for many patients includes endoscopic biopsy surveillance for the early detection of cancer. A fundamental question is the distinction between neoplastic Barrett's epithelium and the surrounding normal tissues of the upper gastrointestinal tract. For example, although it arises in the esophagus, Barrett's epithelium more closely resembles the epithelium of the duodenum at the cytological level. Therefore we are interested to compare the transcriptional profile of Barrett's epithelium to those of normal upper gastrointestinal tissues including gastric epithelium, squamous epithelium of the esophagus and duodenal epithelium. Endoscopic biopsies from each tissue were collected from a series of patients during routine surveillance. Poly A^+ RNA was prepared from pooled samples (2-4 patients/pool) of Barrett's epithelium (4 pools), esophageal squamous epithelium (4 pools), gastric (3 pools) and duodenum (3 pools). Each poly A^+ sample was used to prepare double-stranded cDNA with a T7 promoter. Subsequently fluorescently labeled cRNA, generated by in vitro transcription (IVT) of the cDNA template, was used to interrogate Affymetrix HU6800 and FL6800 chips.

There are three basic questions we would like to address in our analysis:

- Which normal gastrointestinal tissues (squamous epithelium, duodenum epithelium or gastric epithelium) is the neoplastic Barrett's epithelium most similar to?
- Are there any tissue specific gene clusters?
- What are the genes that make Barrett's epithelium similar (or dissimilar) to each of the normal gastrointestinal tissue?

Since we used both Affymetrix HU6800 and FL6800 chips in our experiments, there are some normalization issues for combining the data. The data pre-processing issues will be discussed in Section 2. The similarity analysis will be discussed in Section 3. Section 4 will discuss the cluster analysis on this data set. Finally, a novel algorithm that identifies genes that are responsible for the similarity (or dissimilarity) between tissue samples will be presented in Section 5.

2 Experiments and Data Sets

2.1 Details of the Experiments

In our experiments, tissue samples were pooled from two to four patients. There are a total of four separate pools of Barrett's epithelium (BE), four pools of esophageal squamous epithelium (Sq), three

pools of gastric epithelium (GAS) and three pools of duodenum epithelium (DUO). In our first set of experiments, four pools of Barrett's epithelium, four pools of squamous epithelium, 1 pool of gastric epithelium and 1 pool of duodenum epithelium were used to interrogate the Affymetrix Hu6800 chips (a total of 10 experiments). Let us denote the first set of experiments {BE1, BE2, BE3, BE4, Sq1, Sq2, Sq3, Sq4, GAS1, DUO1}. In our second set of experiments, one pool of Barrett's epithelium, one pool of squamous epithelium, two pools of duodenum epithelium and two pools of gastric epithelium were used to interrogate the Affymetrix FL6800 chips (a total of 6 experiments). Let us denote the second set of experiments {BE5, Sq5, GAS2, GAS3, DUO2, DUO3}. The pools of Barrett's epithelium and squamous epithelium used in the second set of experiments (the FL6800 chips) were identical to one of the four pools used in the first set of experiments (the Hu6800 chips). In particular, BE4 and BE5 was derived from the same pool of tissue samples of Barrett's epithelium, and Sq2 and Sq5 was derived from the same pool of tissue samples. Note that each of the two sets of experiments cover all four types of tissue samples.

The Affymetrix Hu6800 and FL6800 chips consist of approximately 7000 genes. The two types of chips consist of the same genes. However, the Hu6800 format divides the 7000 genes into four separate physical chips (namely, A,B,C,D), while the FL6800 format has all the 7000 genes on one physical chip. The probe sets of the two formats are also different. Figure 1 is a cartoon of the data set. The first set of experiments are shown in red, while the second set in black. In the first set of experiments, approximately one quarter of the 7070 genes are on each of the A,B, C, D chips, and the A, B, C, D chips contain the same genes across different experiments. From our experience, the four chips in the Hu6800 format can have very different overall intensities. For example, in experiment BE1, the A chip can be much brighter than the D chip, while in experiment BE2, the D chip is brighter than the A chip. Both experiments BE1 and BE2 are pooled samples of the Barrett's epithelium. Thus, the challenge is that the data from the four separate chips in the Hu6800 format have to be normalized before data analysis on all the 7070 genes can be performed. Our goal is to combine the data from all the 16 experiments (both the Hu6800 and the FL6800 formats).

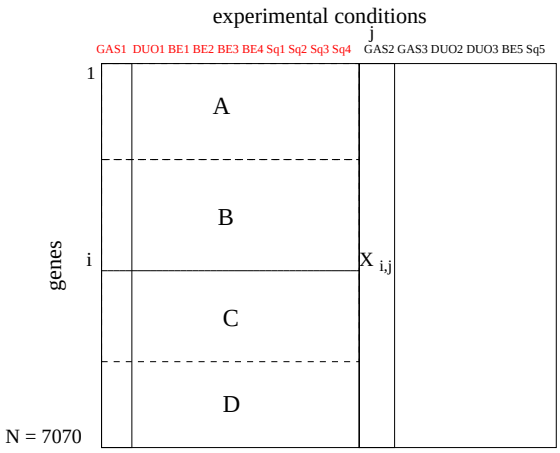

Figure 1: The Barrett's esophagus data set

2.2 Data normalization

The goal of this pre-processing step is to normalize the expression levels of the genes on separate A, B, C, D chips in the Affymetrix Hu6800 chips so as to perform data analysis on all the genes. Let $X_{i,j}$

denote the raw expression level (before normalization) of gene i under experiment j. To motivate the importance of this data normalization step, let us consider the following scenario. Let experiments E_1 and E_2 be experiments from the Hu6800 format. Suppose the overall expression levels of chip A in experiment E_1 are much higher than chip D in the same experiment E_1. Let gene g_i be a gene on chip A, and gene g_j be a gene on chip D, and suppose further that $X_{g_i,E_1} > X_{g_j,E_1}$. However, in experiment E_2, which is the same tissue type as experiment E_1, the overall expression levels of chip A is much lower than that of chip D under experiment E_2, and $X_{g_i,E_2} < X_{g_j,E_2}$. A discrepancy is observed: under the same tissue type, gene g_i is higher expressed than gene g_i under one experiment but not another. Without normalization, we cannot decide if the discrepancy is an artifact of chips A and D having different signal intensities under the two experiments E_1 and E_2, or a result of heterogeneity of tissue samples in the two experiments.

One difficulty of normalization is that the sets of genes on the four separate chips are mostly disjoint. There are only a few control genes that are in common among the four chips. We cannot obtain robust estimates of the mean and the standard deviation of chip intensity with only a few control genes.

Our normalization approach: The basic idea of our normalization approach is to use the data on the FL6800 format to determine the relative intensities of genes on each of the A,B, C, D chips in order to compare the expression levels of genes on different chips under the same experiment. The distributions of the raw expression levels $X_{i,j}$ from each experiment are highly skewed and have a very long tail. An example is shown in Figure 2, which is the histogram of the distribution of the expression levels in experiment Sq2. In the first step of normalization, we took the logarithm of all the expression levels from all 16 experiments. After the log transform, the distribution of the expression levels more closely resembles the normal distribution. The distribution of the log of the expression levels in experiment Sq2 is shown in Figure 3. Then, we normalized the log-transformed expression levels of each of the six experiments from the FL6800 format to mean 50 and standard deviation 10 (the choice of 50 and 10 is arbitrary). For each of the second set of experiments E from the FL6800 format (where E=BE5, Sq5, DUO2, DUO3, GAS2 or GAS3), the average expression levels μ_E^{chip} and standard deviations σ_E^{chip} (where $chip = A, B, C, D$) of corresponding genes on the A, B, C, D chips are computed. The expression levels of the first set of experiments were normalized to have the corresponding mean μ_E^{chip} and standard deviations σ_E^{chip} of the same tissue type. For example, the expression levels of genes in chip A from experiments Sq1, Sq2, Sq3 and Sq4 were scaled to have mean μ_{Sq5}^A and standard deviation σ_{Sq5}^A. The final distribution of experiment Sq2 is shown in Figure 4. In the case of the duodenum epithelium, two experiments (DUO2, DUO3) were done on the FL6800 chips. The average of μ_{DUO2}^{chip} and μ_{DUO3}^{chip}, and the average of σ_{DUO2}^{chip} and σ_{DUO3}^{chip} (where $chip = A, B, C, D$) were used to normalize experiment DUO1. Similarly, GAS1 was normalized with the averages of GAS2 and GAS3.

After this normalization, we can compare expression levels of genes across different chips from the first set of experiments. In terms of our motivating scenario, we can now compare the expression level of gene g_i on chip A to that of gene g_j on chip D. The disadvantage of this approach is that even the same type of tissue samples can be heterogeneous, especially for the neoplastic Barrett's epithelium. This normalized data set is used in all of the analysis described in this technical report.

An alternative normalization approach: Our approach to normalization only applies to situations in which the second set of experiments using the FL6800 chips covers all types of tissue samples. We also experimented with an alternative normalization approach in which the data on the Hu6800 chips is normalized without using the data on the FL6800 chips. The basic idea of this approach is that the average intensity and standard deviation of each of the A,B,C,D chips are scaled to be the same

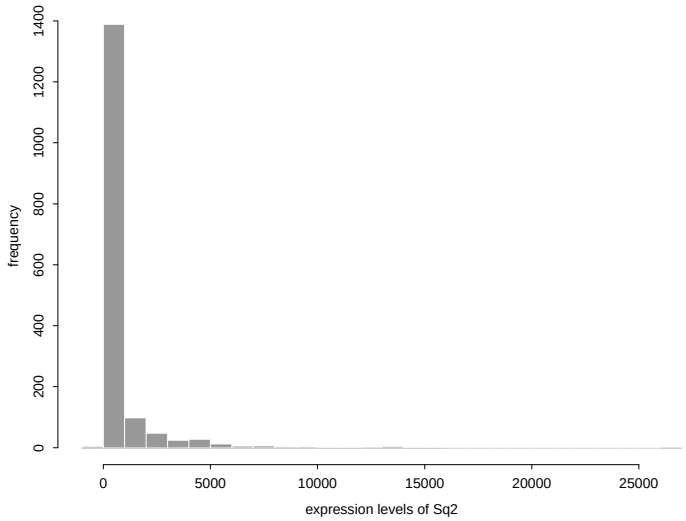

Figure 2: Histogram of the distribution of the expression levels in Sq2

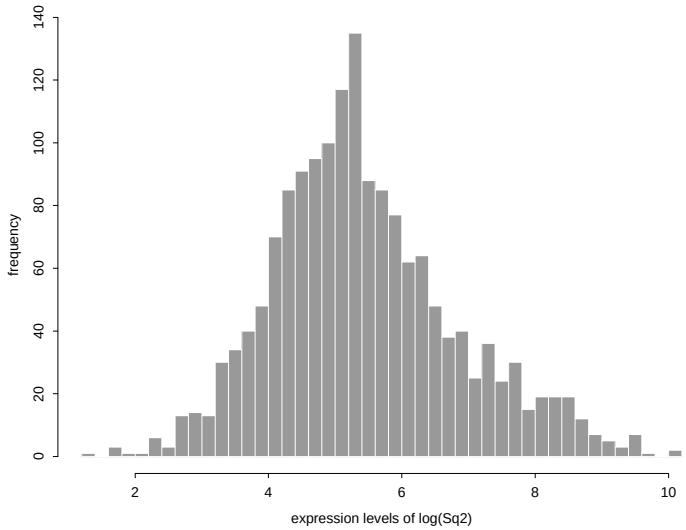

Figure 3: Histogram of the distribution of the expression levels after log transform in Sq2

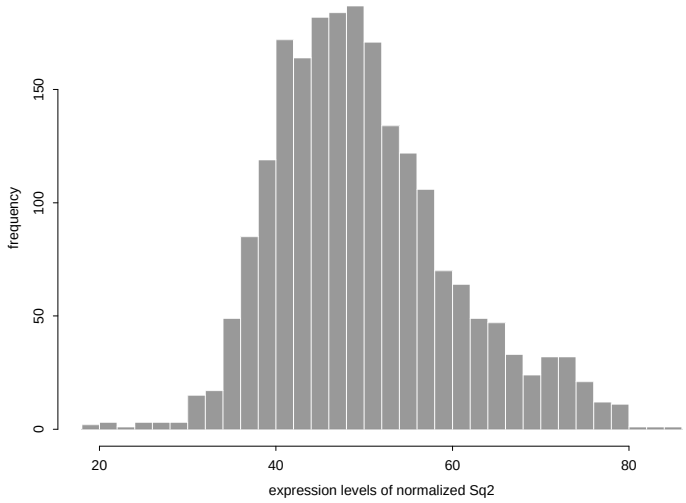

Figure 4: Histogram of the distribution of the expression levels after normalization in Sq2

across different experiments in the Hu6800 format so as to compare expression levels across different experiments. Again, we took the logarithm of all the expression levels from all the experiments. Genes within the same A,B, C or D chips from experiments under the Hu6800 formats were normalized to have the same mean (0) and standard deviation (1). Genes from the FL6800 format were normalized to have the same mean (0) and standard deviation (1) across the 7000 genes. After this normalization, genes in the same chip (A, B, C, or D) have approximately the same distribution as genes in the same respective chip across different experiments. With respect to our motivating scenario, we can now compare the expression levels of genes in chip A under experiment E_1 and the expression levels of genes on chip A under E_2. However, we still cannot compare the expression levels of genes across different chips (A,B, C or D) in the first set of experiments. In terms of our scenario, we still cannot compare the expression levels of genes in chip A and those in chip D. This is because we did not scale the relative intensities across different chips.

This alternative approach implicitly assumes that the average log-transformed expression levels on each of the A,B,C, D chips are comparable. Using this alternative normalized data set in the similarity analysis leads to similar conclusions as the data set normalized with the FL6800 chips.

3 Similarity between tissue samples

One of our fundamental questions is the distinction between neoplastic Barrett's epithelium and the surrounding normal tissues of the upper gastrointestinal tract. We used the Pearson's correlation coefficient [Pearson, 1896] to compare the pairwise similarities between tissue samples. Using the notations in Figure 1. The similarity (Pearson correlation coefficient) between experiment j and experiment k is

$$\frac{\sum_{g=1}^{N}(X_{g,j}-\mu_j)*(X_{g,k}-\mu_k)}{\sqrt{\sum_{g=1}^{N}(X_{g,j}-\mu_j)^2*\sum_{g=1}^{N}(X_{g,k}-\mu_k)^2}} \tag{1}$$

where $\mu_j = \frac{\sum_{g=1}^{N} X_{g,j}}{N}$. The normalized data is used to compute the correlation coefficients. The theory of Pearson's correlation has an implicit assumption on normality of the data. The distribution of normalized data from approach 1 resembles the normal distribution (an example is shown in Figure 4).

Since we have multiple experiments on each tissue type, we averaged the normalized expression levels across experiments with the same tissue type in the same set of experiments in order to summarize the similarities between different tissue types. Then, the Pearson's correlation coefficient was applied to each pair of tissue types and in each set of experiments. The results are shown in Table 1. Due to the different probe sets used by the Hu6800 and the FL6800 formats in the two sets of experiments, we restrict our comparison within the same set of experiments. In Table 1, the notation E(r-s) (where $r < s$) means that the expression levels in experiments $Er, \ldots, Es$ were averaged. For example, BE(1-4) represents the expression levels in experiments BE1, BE2, BE3 and BE4 were averaged. The first set of experiments is shown in red. Table 1 shows the point estimates of the correlation coefficients. We also computed the 95% confidence intervals for the correlation coefficients using Fisher's transform [Snedecor and Cochran, 1980] (not shown here).

	GAS1	DUO1	BE(1-4)	Sq(1-4)	GAS(2-3)	DUO(2-3)	BE5	Sq5
GAS1	1.000	0.807	0.851	0.751	0.864	0.763	0.805	0.741
DUO1	0.807	1.000	0.841	0.732	0.761	0.872	0.792	0.719
BE(1-4)	0.851	0.841	1.000	0.830	0.810	0.782	0.865	0.795
Sq(1-4)	0.751	0.732	0.830	1.000	0.732	0.689	0.729	0.892
GAS(2-3)	0.864	0.761	0.810	0.732	1.000	0.861	0.863	0.777
DUO(2-3)	0.763	0.872	0.782	0.689	0.861	1.000	0.872	0.748
BE5	0.805	0.792	0.865	0.729	0.863	0.872	1.000	0.796
Sq5	0.741	0.719	0.795	0.892	0.777	0.748	0.796	1.000

Table 1: Average correlation coefficients between tissue types in the same set of experiments.

Let $S(x, y)$ be the pairwise similarity between experiment x and experiment y. From Table 1, $S(GAS1, DUO1) = 0.807$, $S(GAS1, Sq(1-4)) = 0.751$ and $S(DUO1, Sq(1-4)) = 0.732$. Therefore, the gastric epithelium and the duodenum epithelium are more similar to each other than to the squamous epithelium (because $S(GAS1, DUO1) > S(GAS1, Sq(1-4))$ and $S(GAS1, DUO1) > S(DUO1, Sq(1-4))$). Even the low end of the confidence interval for $S(GAS1, DUO1)$ is greater than the high end of the confidence interval for $S(GAS1, Sq(1-4))$ and $S(DUO1, Sq(1-4))$. Similarly, the correlation coefficients in the second set of experiments support the same conclusion. The comparison of the expression profiles of the three normal gastrointestinal tissues is consistent with the more similar morphology and physiological role (secretory) of gastric and duodenal epithelia, when compared to the different morphology of non-secretory esophageal squamous epithelium.

For the first set of experiments, the point estimates of the correlation coefficients between the Barrett's epithelium and each of the gastric epithelium, duodenum epithelium and squamous epithelium are comparable. The confidence intervals for the correlation coefficients also overlap. However, for the second set of experiments, the Barrett's epithelium, BE5, is more similar to the gastric epithelium, GAS(2-3), and the duodenum epithelium, DUO(2-3), than to the squamous epithelium, Sq5. It turns out that this discrepancy is due to the heterogeneity of the neoplastic Barrett's epithelium. Table 2 in the Appendix shows the point estimates of the correlation coefficients between all 16 experiments without averaging the expression levels over the same tissue type. Again, the first set of experiments are shown in red, while the second set is shown in black. (We also computed the 95% confidence interval, but the results are not shown here). From Table 2, we can see that experiment BE1 from the first set of

experiments has lower similarity to the gastric epithelium (GAS1) than to the squamous epithelium (Sq1, Sq2, Sq3, Sq4). On the other hand, experiment BE4 (also from the first set of experiments) has higher similarity to the gastric epithelium (GAS1) than to the squamous epithelium (Sq1, Sq2, Sq3, Sq4). In the second set of experiments, experiment BE5 shows the same relative similarities as BE4, *i.e.,* $S(BE5, Sq5) < S(BE5, GAS2)$ and $S(BE5, Sq5) < S(BE5, GAS3)$. In fact, experiments BE4 and BE5 used the same pooled tissue sample, but they are interrogated to the Hu6800 and FL6800 format respectively. Therefore, the discrepancy we observed using the average expression levels across tissue types in Table 1 merely reflects the heterogeneity of the neoplastic Barrett's epithelium.

From Table 2, experiment BE5 is most similar to experiment BE4 across all the experiments, even though BE4 and BE5 were interrogated to different chip formats. Similarly, experiment Sq5 is most similar to experiment Sq2 across all the experiments. This shows that our normalized data and similarity comparisons are robust because experiments BE4 and BE5, Sq2 and Sq5 used the same pooled tissue samples.

4 Cluster Analysis

In order to identify tissue specific clusters of genes, the entire normalized data set is filtered to focus on genes that are differentially expressed in different tissue types. After we determined a set of differentially expressed genes, we need to choose a clustering algorithm. Finally, we applied the chosen clustering algorithm to obtain tissue specific clusters.

4.1 Filtering

Our procedure to identify genes that are differentially expressed in different tissue types is similar to the standard procedure of the analysis of variance (ANOVA) [Zar, 1984]. Suppose we have independent samples from each of the k different populations, and the sample size from population i is n_i (where $i = 1, 2, \ldots, k$). Let $Y_{i,j}$ be an expression level from population i, where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$. In the standard ANOVA procedure, $Y_{i,j}$'s are assumed to be independent, normal, $E[Y_{i,j}] = \mu_i$, $Var[Y_{i,j}] = \sigma^2$, and the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ versus $H_1 : H_0$ *is false* is tested. Note that the population variances are assumed to be equal. Let $n = n_\cdot = \sum_i n_i$, $Y_{i\cdot} = \sum_j Y_{ij}/n_i$, $Y_{\cdot\cdot} = \sum_i Y_{i\cdot}/\sum_i n_i = \sum_i \sum_j Y_{ij}/n$. The test statistic in the standard ANOVA procedure is the ratio of the between-population mean square to the residual mean square, *i.e.,*

$$\frac{\sum_i n_i (Y_{i\cdot} - Y_{\cdot\cdot})^2}{k-1} \Big/ \frac{\sum_i \sum_j (Y_{ij} - Y_{i\cdot})^2}{n-k} \tag{2}$$

which follows the F-distribution with (k-1, n-k) degrees of freedom.

For each gene, we tested the null hypothesis $H_0 : \mu_{BE} = \mu_{Sq} = \mu_{GAS} = \mu_{DUO}$ versus $H_1 : H_0$ *is false*. A gene is said to be *differentially expressed* if the null hypothesis H_0 is rejected. There are four tissue types in our experiments: the Barrett's epithelium, gastric epithelium, duodenum epithelium and squamous epithelium, *i.e.,* k is 4. The sizes of the tissue types, n_i are 5, 5, 3, and 3 for the Barrett's, squamous, gastric and duodenum epithelium respectively.

Our idea is to use the test statistic in Equation 2, but instead of assuming that the test statistic follows the F-distribution with (k-1, n-k) degrees of freedom, an empirical distribution for the test statistic is computed. Due to the small sample sizes (3 or 5), the assumption of the F distribution can potentially have a large impact on the hypothesis testing. In the derivation of the test statistic, the normality assumption is used to show that the distribution of the test statistic in Equation 2 follows the

F-distribution. Therefore, by generating an empirical distribution to compute the significance level, our approach does *not* assume the normality of the expression levels $Y_{i,j}$'s from each tissue type.

An empirical distribution for each gene is simulated by randomly permuting the expression levels of that gene from all the experiments, and by repeating the random permutation many many times (3000 times in our implementation). If the test statistic of Equation 2 from the empirical distribution of a gene g is greater than the observed test statistic from the data less than 5% in all the random trials, then we reject the null hypothesis H_0 at the 0.05 significance level. The gene g passes the filter, and is considered in the cluster analysis.

For a fixed k (=4) and a fixed n (=16), the test statistic in Equation 2 is equivalent to the ratio of the between tissue type mean square to the residual mean square. Since our goal is to identify genes that are differentially expressed in different tissue types, a large ratio of the between tissue type mean square to the residual mean square is preferred. Intuitively, our empirical testing procedure determines whether the observed ratio from the data is large enough so that it is not easily obtained by chance.

We applied the above modified ANOVA procedure to the thresholded normalized data set from approach 2. Data values with very low expression levels that are marked with low confidence by the Affymetrix software were not considered in the normalization step. After normalizing the data with approach 2, we thresholded all the low confidence data with a value that is slightly lower than the lowest expression level marked with confidence. For 1095 genes (out of 7070 genes), the equal population mean null hypothesis is rejected at the 0.05 significance level, and hence passing the filter.

4.2 Choosing a clustering algorithm

With the filtered data set, the next problem is to choose a clustering algorithm for the data. We used the *figure of merit* methodology in [Yeung *et al.*, 2000] to compare the performance of different clustering algorithms. The basic idea of the figure of merit (FOM) methodology is to apply a clustering algorithm to the data from all but one experiment. The remaining experiment is used to assess the predictive power of the resulting clusters—meaningful clusters should exhibit less variation in the remaining experiment than clusters formed by chance. The predictive power of the resulting clusters is measured by the within-cluster variance, and is called the *figure of merit* (FOM). A clustering result with a small FOM implies low within-cluster variance, which in turn is an indication of high predictive power. The definition of FOM does not allow direct comparisons over different numbers of clusters. Therefore, the FOM is plotted against the number of clusters in typical FOM analyses.

Figure 5 shows the result of applying the FOM methodology to the filtered Barrett's esophagus data (1095 genes). Correlation coefficient was used to compute pairwise similarities of genes. Three hierarchical clustering algorithms [Jain and Dubes, 1988] (average-link, single-link, complete-link), two partitional algorithms (k-means [Jain and Dubes, 1988], and Cluster Affinity Search Technique (CAST) [Ben-Dor and Yakhini, 1999]), and the random algorithm were compared. The random algorithm is a benchmark in which all genes are randomly assigned to clusters. A good clustering algorithm should do much better than the random algorithm. In our implementation, k-means is initialized with the results from hierarchical average-link. From Figure 5, the single-link algorithm achieves only slightly lower FOM than the random algorithm, which means that the performance of single-link is not satisfactory. The k-means and CAST algorithms achieve the lowest FOM, and have comparable performance. The FOM declines drastically up to around 8 clusters, so the number of clusters is estimated to be approximately 8.

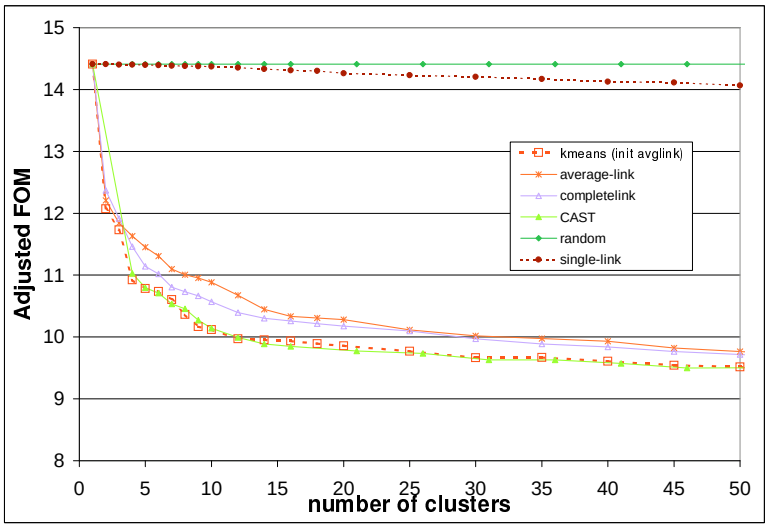

Figure 5: FOM analysis on the Barrett's esophagus data

4.3 Tissue specific clusters

From the FOM analysis, we applied the CAST algorithm to the filtered Barrett's esophagus data set (1095 genes) to obtain 8 clusters. Tissue specific clusters were obtained. For example, Figure 6 shows the expression profile of a Barrett specific cluster, and Figure 7 shows a squamous specific cluster. In Figures 6 and 7, the x-axis represents all the 16 experiments, and the y-axis shows the normalized expression levels. The solid line represents the average expression level in each experiment, and the dotted lines show one standard deviation above and below the average expression level in each cluster. From the figures, genes in the Barrett specific cluster (Figure 6) show relatively high expression levels in the five experiments using the Barrett's epithelium tissue, while genes in the squamous specific cluster (Figure 7) show relatively high expression levels in the five experiments using the squamous epithelium tissue sample. Many interesting genes were found from these tissue specific clusters. The Barrett specific cluster included genes associated with cell cycle progression (P1cdc47, PCM-1), cell migration (urokinase-type plasminogen receptor), growth regulation (TGF-beta superfamily, amphiregulin, Cyr61) stress responses (calcyclin, ATF3, TR3 orphan receptor) as well as epithelial cell surface antigens (epsilon-BP, Human surface antigen, integrin beta 4). The squamous specific cluster included oncogenes (pim-1, met, P47 LBC), a number of proteinase inhibitors (maspin, elafin, monocyte/neutrophil elastase inhibitor, cystatin M, cystatin B, squamous cell carcinoma antigen, urokinase inhibitor), proteases (protease M, calcium dependent protease) and a series of small proline rich proteins (sprI, sprII, SPRR2B, SPR2-1, SPRR1A) implicated in various cellular stress responses. For more detailed biological interpretation, please refer to our paper [Barrett *et al.*, 2000].

4.4 Discussion

A careful inspection of the clusters in Figures 6 and 7 shows that the experiments using the same pool of tissue samples (BE4 and BE5, Sq2 and Sq5) do not have identical normalized expression levels. The differences between the normalized expression levels of the same tissue samples hybridized to both HU6800 and FL6800 chips reflect the experimental variation in using either the same cDNA (BE4 and BE5) or the same poly A+ (Sq2 and Sq5) as starting material to generate the separate pools

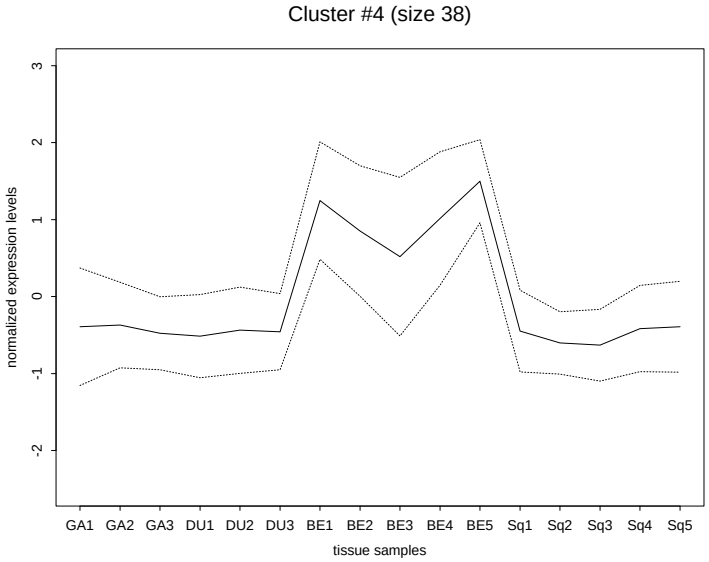

Figure 6: Barrett specific cluster

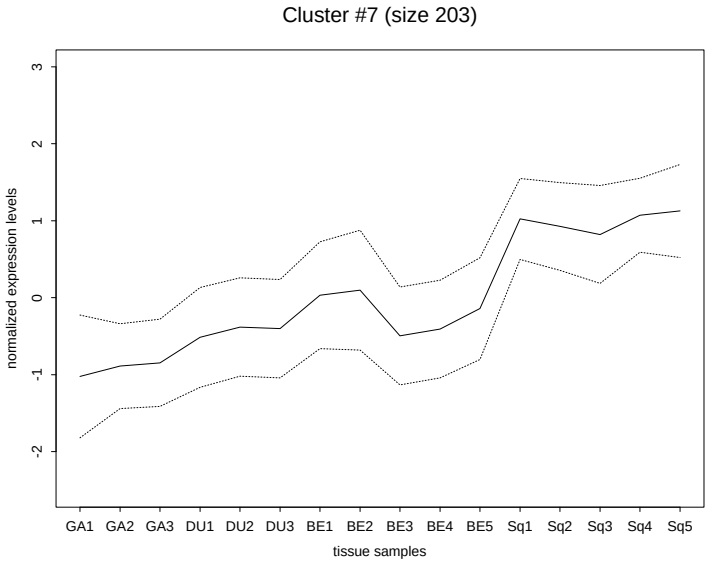

Figure 7: squamous specific cluster

of cRNA for each hybridization.

In our filtering procedure, the expression levels of the experiments corresponding to each tissue type were assumed to be independent. This is not the case for the experiments with the same pool of tissue samples (BE4 and BE5, Sq2 and Sq5). It would be interesting to modify our current approach to account for the dependence between tissue samples.

5 Genes driving the similarity between tissue samples

In the calculation of the pairwise similarities between tissue samples in Section 3, the expression levels of all the 7070 genes were taken into consideration. One interesting question would be to determine the genes that drive the similarity between tissue types. For example, what are the genes that make the Barrett's epithelium (BE) similar to the squamous epithelium (Sq)? What are the genes that make the Barrett's epithelium (BE) different from the squamous epithelium (Sq)?

In order to answer this question, we developed a novel algorithm, the GENEEXTRACT algorithm. This algorithm is motivated by the ideas behind the Kendall's coefficient of concordance[Kendall, 1970], which is a measure of rank association.

5.1 Kendall's coefficient of concordance

In this subsection, the Kendall's coefficient of concordance and the necessary notations are introduced. Using the notations in Figure 1, let $X_{i,j}$ be the expression level of gene i under experiment j, where $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, m$. In the case of the Barrett's esophagus data, N = 7070 and m = 16.

Definition 1 *A pair of genes g_i and g_j are concordant with respect to experiments E_1 and E_2 if* $(X_{g_j,E_1} - X_{g_i,E_1}) * (X_{g_j,E_2} - X_{g_i,E_2}) > 0$.

Definition 2 *A pair of genes g_i and g_j are discordant with respect to experiments E_1 and E_2 if* $(X_{g_j,E_1} - X_{g_i,E_1}) * (X_{g_j,E_2} - X_{g_i,E_2}) < 0$.

In other words, if genes g_i and g_j are *concordant*, either gene g_i has higher expression levels than gene g_j in both experiments or gene g_i has lower expression levels in both experiments. For a pair of *discordant* genes, their expression levels go up or down in opposite direction in both experiments. Note that the magnitudes of expression levels are not necessary to determine if a pair of genes is concordant or discordant, only the relative *ranks* are necessary. The rank of an object is the relative position of a set of objects if all the objects are arranged in increasing order of a given measure. Let $R(g, E)$ be the rank of a gene g in experiment E. It is clear that the following three conditions are true:

- $R(g_i, E) > R(g_j, E)$ if and only if $X_{g_i,E} > X_{g_j,E}$.
- $R(g_i, E) < R(g_j, E)$ if and only if $X_{g_i,E} < X_{g_j,E}$.
- $R(g_i, E) = R(g_j, E)$ if and only if $X_{g_i,E} = X_{g_j,E}$.

Therefore, the magnitudes of expression levels of genes in Definition 1 and Definition 2 can be replaced by the relative ranks of genes.

Example 1 *Table 3 shows the relative ranks of expression levels of genes G1 to G8 with respect to two experiments E_1 and E_2. For example, G2 has the lowest expression level in experiment E_1 because it has rank 1, while G1 has the lowest expression level in experiment E_2.*

In this example, genes (G7, G8) are concordant because $(4-7)*(3-4) > 0$. Similarly, genes (G3, G4) are discordant because $(2-6)*(6-2) < 0$.

	Experiment E_1	*Experiment E_2*
G1	5	1
G2	1	8
G3	2	6
G4	6	2
G5	3	5
G6	8	7
G7	4	3
G8	7	4

Table 3: An example showing Kendall's coefficient of concordance.

Let C be the number of pairs of genes that are concordant, and D be the number of discordant pairs of genes. Suppose there are a total of N objects (genes in our case). The Kendall's coefficient of concordance (τ) is defined to be the difference of the number of concordance pairs and the number of discordant pairs divided by the total number of possible pairs, *i.e.,*

$$\tau = \frac{C - D}{\binom{N}{2}} \tag{3}$$

From the definition in Equation 3, it is clear that τ lies between -1 and 1 ($\tau = 1$ when $D = 0$, and $\tau = -1$ when $C = 0$). When the number of concordant pairs is equal to the number of discordant pairs (*i.e.*, $C = D$), $\tau = 0$ which has the interpretation that the two experiments are uncorrelated. In Example 1, C = 10, D = 18, and N = 28, so the Kendall's coefficient of concordance for experiments E_1 and E_2, τ, is -0.286.

In Example 1, the ranks of the genes in each experiment are distinct, *i.e.,* there are no *ties*. In general, for any pair of genes, they must be concordant or discordant or tied. In the numerator in the formula for the Kendall's coefficient of concordance (Equation 3), ties are not considered. Therefore, in the case of ties, the denominator in Equation 3 has to be adjusted. Specifically, the number of pairs of tied pairs from each experiment has to be subtracted from the denominator.

5.2 Reduction to a graph problem (max-clique)

From the definitions in Section 5.1, it is clear that concordant genes contribute to the similarity of two experiments, while discordant genes contribute to the dissimilarity of two experiments. If we compute the Kendall's coefficient of concordance within a subset of genes that are all concordant with each other, the Kendall's coefficient of concordance will be 1. One of the questions that we would like to address in this expression study is to identify genes that make the Barrett's epithelium distinct from other normal gastrointestinal tissues. Motivated by the concepts of concordance and discordance in Kendall's coefficient of concordance, a subset of genes that are concordant to each

other in two experiments is said to make the two experiments "similar". Similarly, a subset of genes that are discordant to each other in two experiments is said to make the two experiments "dissimilar".

Special case: two experiments Let us first consider the problem of determining a subset of concordant genes in two experiments E_1 and E_2. With the notion of concordant genes, we can identify pairs of genes whose expression levels go up or down in the same direction. The problem with the concordant notion is that it is a *pairwise* concept. In order to find a subset of genes that are all concordant to each other, we can reduce this problem to a graph problem. Let $G = (V, E)$, where V is the set of vertices and E is the set of edges. The graph G has N vertices (each vertex corresponds to a gene). Edge (g_i, g_j) is in the graph if genes g_i and g_j are concordant with respect to experiments E_1 and E_2. We can reduce Example 1 to the graph in Figure 8 with a vertex for each of the eight genes, and an edge for each pair of concordant genes.

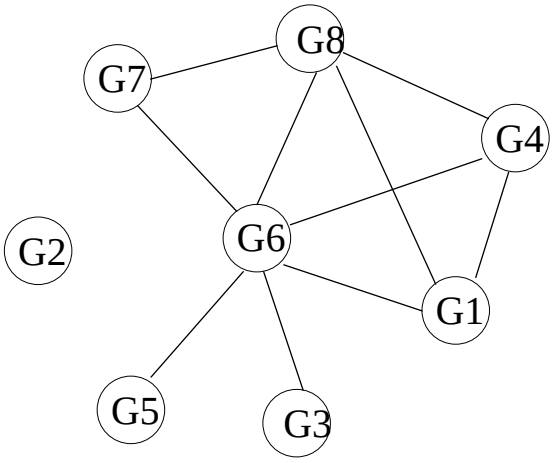

Figure 8: Example 1 is reduced to a graph problem.

After reducing our problem to a graph, the next step is to find a subset of genes (or vertices in the graph) that are all concordant, *i.e.*, to find the largest subset of genes that are all connected. In the computer science literature, this is known as the *max-clique* problem. A subgraph is a subset of vertices and edges in a graph. A *clique* is a complete subgraph, *i.e.*, each vertex in a clique is connected to every other vertex in the clique. A max-clique is a clique with the maximum number of vertices. In Example 1 above, G1, G4, G6, G8 forms a max-clique. For our problem, the max-clique is a subset of maximum number of genes (vertices) such that they are all concordant (or discordant) to each other in the subset.

The max-clique problem is shown to be NP-complete and is known to be a very difficult problem even to approximate [Hastad, 1996]. However, there are many approximation heuristics that can be used to solve the max-clique problem. We used the *reactive local search* (RLS) implementation developed by [Battiti and Protasi, 2000]. The basic idea of the RLS algorithm is that it is a local search algorithm with techniques to prohibit moves that would create cycles in the search trajectory and to exploit new parts of the total search space. The RLS implementation achieved significantly better results than all other max-clique algorithms at the DIMACS implementation challenge [Johnson and Trick, 1996]. Moreover, their implementation is easy to use and is available on the world wide web.

For the problem of finding a subset of genes that are all discordant to each other in two experiments, an edge (g_i, g_j) is added to the graph $G = (V, E)$ if genes g_i and g_j are discordant with respect to the

two experiments. The max-clique on the graph G with discordant edges represents the largest subset of genes that are all discordant to each other in the subset.

The reduction to max-clique allows us to find a subset of genes that are either all concordant or all discordant to each other. If we allow weights on the edges of the graph, we can represent both concordant and discordant gene pairs in the same graph. Let $G^{w1} = (V, E^{w1})$ be a weighted graph, and w_{ij} be the weight on edge (g_i, g_j), where $i, j = 1, 2, \ldots, N$. If genes g_i and g_j are concordant with respect to experiments E_1 and E_2, weight w_{ij} is 1. On the other hand, if genes g_i and g_j are discordant, weight w_{ij} is -1.

Theorem 1 *Ignoring ties, finding the largest subgraph $G' = (V', E')$ in G^{w1} such that $F_{E_1,E_2} = \frac{\sum_{(i,j)\in E'} w_{ij}}{|E'|}$ is maximized (or minimized) is equivalent to finding the largest subset of genes for which the Kendall's coefficient of concordance is maximized (or minimized) with respect to experiments E_1 and E_2.*

Proof Outline: The theorem follows directly from the following two observations. With weights on concordant pairs of genes being 1 and weights on discordant pairs of genes being -1, it is easy to see that $\sum_{(i,j)\in E'} w_{ij} = C - D$, where C and D are the number of pairs of concordant and discordant genes in the subgraph G'. If there are no ties, G^{w1} is a complete graph with $\binom{N}{2}$ edges. □

The formulation in Theorem 1 establishes the reduction of our problem of finding a subset of genes with the maximum (or minimum) Kendall's coefficient of concordance to a weighted graph G^{w1}. The following two corollaries shows that the formulation in Theorem 1 is equivalent to the max-clique formulation.

Corollary 1 *Let $G = (V, E)$ be a graph such that the set of vertices are the set of genes. Let $g_i, g_j \in V$. Edge $(g_i, g_j) \in E$ if genes g_i and g_j are concordant with respect to experiments E_1 and E_2. Let $G^{w1} = (V, E^{w1})$ be another graph with the same set of vertices as G, but the edges are weighted. Edge $(g_i, g_j) \in E^{w1}$ has weight w_{ij} 1 if genes g_i and g_j are concordant with respect to experiments E_1 and E_2, and has weight -1 if genes g_i and g_j are discordant. Finding the largest subgraph $G' = (V', E')$ in G^{w1} such that $F_{E_1,E_2} = \frac{\sum_{(i,j)\in E'} w_{ij}}{|E'|}$ is maximized is equivalent to finding the max-clique in G.*

Proof Outline: The maximum value of F_{E_1,E_2} is 1, which can only be obtained by having $w_{ij} = 1$ for all edge $(i, j) \in E'$, which in turn implies that only concordant gene pairs are allowed in the subgraph G'. It follows that the largest subgraph G' in G^{w1} such that F is maximized is the same as the max-clique in G. □

Corollary 2 *Let $G = (V, E)$ be a graph such that the set of vertices are the set of genes. Let $g_i, g_j \in V$. Edge $(g_i, g_j) \in E$ if genes g_i and g_j are discordant with respect to experiments E_1 and E_2. Let $G^{w1} = (V, E^{w1})$ be another graph with the same set of vertices as G, but the edges are weighted. Edge $(g_i, g_j) \in E^{w1}$ has weight w_{ij} 1 if genes g_i and g_j are concordant with respect to experiments E_1 and E_2, and has weight -1 if genes g_i and g_j are discordant. Finding the largest subgraph $G' = (V', E')$ in G^{w1} such that $F_{E_1,E_2} = \frac{\sum_{(i,j)\in E'} w_{ij}}{|E'|}$ is minimized is equivalent to finding the max-clique in G.*

The proof of Corollary 2 is very similar to that of Corollary 1, and so is not shown here. From Theorem 1, Corollaries 1 and 2, the problem of finding a subgraph in which the Kendall's coefficient

of concordance is maximized (or minimized) is equivalent to finding the max-clique for concordant (or discordant) pairs of genes.

More general case: more than one experiment for each tissue type In Theorem 1, edges either have weight 1 or -1 depending on whether a pair of genes are concordant or discordant with respect to *two* experiments. In the case of the Barrett's esophagus data set, our goal is to identify genes driving the similarity between different tissue types, and there are more than one experiment for each tissue type, *i.e.,* there are three experiments for each of the duodenum epithelium and gastric epithelium, and five experiments for each of the Barrett's epithelium and squamous epithelium. One natural measure of the similarity of two tissue types over various experiments is the average similarity between all pairs of experiments from each tissue type. In this formulation with multiple experiments in each tissue type, the edges in the graph have weights other than 1 and -1.

Let T_1 and T_2 be different tissue types. Let $\{E_1^1, E_2^1, \ldots, E_p^1\}$ be experiments done on tissue type T_1, and $\{E_1^2, E_2^2, \ldots, E_q^2\}$ be experiments done on tissue type T_2. Let $G^w = (V, E^w)$ be a graph with genes as the set of vertices. Edge (g_i, g_j) has weight $w_{ij}^{tot} = \frac{k}{pq}$, where k is the number of pairs of experiments in which genes g_i and g_j are concordant minus the number of pairs of experiments over the two tissue types in which genes g_i and g_j are discordant. Let $C_{E_r^1, E_s^2}$ and $D_{E_r^1, E_s^2}$ be the number of concordant and discordant pairs of genes with respect to experiments E_r^1 and E_s^2, where $r = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, q$. $w_{ij}^{tot} = \frac{k}{pq}$ can be rewritten as $\frac{\sum_{r=1}^{p}\sum_{s=1}^{q}(C_{E_r^1,E_s^2} - D_{E_r^1,E_s^2})}{pq}$. Let $G' = (V', E')$ be a subgraph of G^w. Define $F^{tot} = \frac{\sum_{(i,j)\in E'} w_{ij}^{tot}}{|E'|}$.

Theorem 2 *Ignoring ties, the problem of finding a subset of genes such that the average Kendall's coefficient of concordance over multiple experiments from each tissue type is maximized (or minimized) is equivalent to finding a subgraph G' in G^w in which F^{tot} is maximized (or minimized).*

Proof: Let us consider a pair of experiments over the two tissue types, E_r^1 and E_s^2, where $r = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, q$. The Kendall's coefficient of concordance between this pair of experiments, $\tau_{E_r^1,E_s^2}$, is $\frac{C_{E_r^1,E_s^2} - D_{E_r^1,E_s^2}}{\binom{N}{2}}$, where $C_{E_r^1,E_s^2}$ and $D_{E_r^1,E_s^2}$ are the number of pairs of concordant and discordant genes with respect to experiments E_r^1 and E_s^2 respectively. The average Kendall's coefficient of concordance over all pairs of experiments from each tissue type, $\bar{\tau}$, is $\frac{\sum_{r=1}^{p}\sum_{s=1}^{q}\tau_{E_r^1,E_s^2}}{pq}$.

$$
\begin{aligned}
F^{tot} &= \frac{\sum_{(i,j)\in E'} w_{ij}^{tot}}{|E'|} \\
&= \frac{1}{pq} * \frac{\sum_{(i,j)\in E'}\sum_{r=1}^{p}\sum_{s=1}^{q}(C_{E_r^1,E_s^2} - D_{E_r^1,E_s^2})}{|E'|} \\
&= \frac{\sum_{r=1}^{p}\sum_{s=1}^{q}\tau_{E_r^1,E_s^2}}{pq} \\
&= \bar{\tau} \quad \square
\end{aligned}
$$

5.3 The GeneExtract Algorithm

In order to find the largest subset of genes that are concordant (or discordant) to each other in the subset over multiple experiments from two different tissue types, we can find the max-clique by considering only edges (g_i, g_j) such that genes g_i and g_j are concordant (or discordant) in *all* pairs of experiments from two different tissue types. This is a very restrictive condition. In terms of the weighted graph G^w

formulation, edge (g_i, g_j) has weight $w_{ij}^{tot} = 1$ if and only if genes g_i and g_j are concordant in *all* pairs of experiments from two different tissue types. It follows that the restrictive formulation described is equivalent to finding a subgraph G' in G^w such that F^{tot} (and hence $\bar{\tau}$ from Theorem 2) is 1. Noise from experiments may make a pair of genes not concordant under one pair of experiments. Consider an example of a weighted graph G^w in Figure 9. There are six vertices (genes), and edges are labeled with their weights w_{ij}^{tot}. For clarity of the figure, edges with weight 1 are shown in red, edges with weight $\frac{3}{4}$ are shown in blue, edges with weight $\frac{1}{4}$ are shown in black, and edges with weight $-\frac{1}{4}$ are shown in dotted lines. If we want to find the largest subgraph G' such that all pairs of genes in the subgraph are concordant to each other over all pairs of experiments from the two tissue types, only edges with weight 1 will be considered, *i.e.*,, the resulting subgraph consists of genes G1, G2 and G3. However, if we relax our formulation and want to find the largest subgraph for which $\bar{\tau}$ is at least 0.8, the resulting subgraph consists of genes G1, G2, G3, G4 and G5. In the case of the Barrett's esophagus data, there are 25 pairs of experiments over the Barrett's epithelium and the squamous epithelium. Restricting our attention to the largest subset of genes for which the genes are discordant with respect to all the 25 pairs of experiments results in only 5 genes (out of a total of 7070 genes). Therefore, our implementation aims to find a subgraph of very "high" or "low" average Kendall's coefficient of concordance, $\bar{\tau}_{given}$ ($\bar{\tau}_{given}$ is specified by the user).

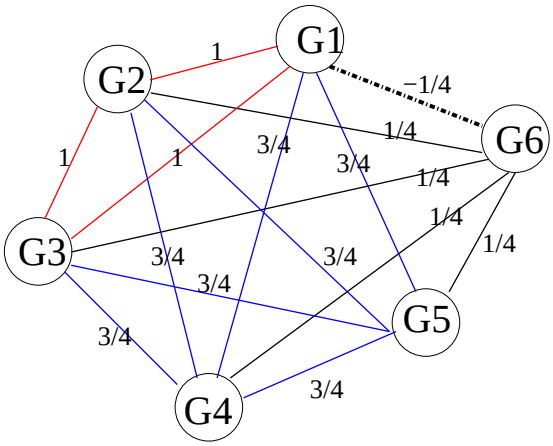

Figure 9: An example illustrating .

In the GeneExtract algorithm, we start by finding a subset of genes for which the genes are concordant (or discordant) to each other with respect to all pairs of experiments from the two tissue types. Then, vertices that are highly connected to the initial subset are greedily added. The details of the algorithm for finding a subset of genes that are highly concordant or discordant are shown below.

GeneExtract algorithm for concordant genes

- Use the unweighted formulation to form a graph $G = (V, E)$: edge $(g_i, g_j) \in E$ if genes g_i and g_j are concordant in all pairs of experiments from the two different tissue types.
- Use the RLS algorithm to find a max-clique, MC, in G.
- Use the weighted formulation in Theorem 2 to form a graph G^w, *i.e.*, edge (g_i, g_j) has weight $w_{ij}^{tot} = \frac{k}{pq}$, where k is the number of pairs of experiments that genes g_i and g_j are concordant minus the number of pairs of experiments over the two tissue types that genes g_i and g_j are discordant.
- For each vertex v not in the max-clique MC, compute total weight $TW(v) = \sum_{u \in MC} w_{uv}$ in G^w.
- The clique MC should have the average Kendall's coefficient of concordance over all pairs of experiments from different tissue types, $\bar{\tau}$, equal to 1.
- Let $MC_{extend} = (V', E')$ be the extended subgraph returned by this algorithm, where $V' \subset V$. Initialize MC_{extend} to be MC. Define $F_{extend}^{tot} = \frac{\sum_{(i,j) \in E'} w_{ij}^{tot}}{|E'|}$. From Theorem 2, F_{extend}^{tot} is equal to the average Kendall's coefficient of concordance of the set of genes in MC_{extend}. Initially, $F_{extend}^{tot} = \bar{\tau} = 1$.
- Repeat until $F_{extend}^{tot} < \bar{\tau}_{given}$,
  - Add vertex v that is not currently in MC_{extend} that has the highest total weight to the original clique $TW(v)$.
  - Recompute F_{extend}^{tot} with the additional vertex v.
- Return MC_{extend} with the smallest F_{extend}^{tot} that exceeds $\bar{\tau}_{given}$.

GeneExtract algorithm for discordant genes

- Use the unweighted formulation to form a graph $G = (V, E)$: edge $(g_i, g_j) \in E$ if genes g_i and g_j are discordant in all pairs of experiments from the two different tissue types.
- Use the RLS algorithm to find a max-clique, MC, in G.
- Use the weighted formulation in Theorem 2 to form a graph G^w, *i.e.*, edge (g_i, g_j) has weight $w_{ij}^{tot} = \frac{k}{pq}$, where k is the number of pairs of experiments that genes g_i and g_j are concordant minus the number of pairs of experiments over the two tissue types that genes g_i and g_j are discordant.
- For each vertex v not in the max-clique MC, compute total weight $TW(v) = \sum_{u \in MC} w_{uv}$ in G^w.
- The clique MC should have the average Kendall's coefficient of concordance over all pairs of experiments from different tissue types, $\bar{\tau}$, equal to -1.
- Let $MC_{extend} = (V', E')$ be the extended subgraph returned by this algorithm, where $V' \subset V$. Initialize MC_{extend} to be MC. Define $F_{extend}^{tot} = \frac{\sum_{(i,j) \in E'} w_{ij}^{tot}}{|E'|}$. From Theorem 2, F_{extend}^{tot} is equal to the average Kendall's coefficient of concordance of the set of genes in MC_{extend}. Initially, $F_{extend}^{tot} = \bar{\tau} = -1$.
- Repeat until $F_{extend}^{tot} > \bar{\tau}_{given}$,
  - Add vertex v that is not currently in MC_{extend} that has the lowest total weight to the original clique $TW(v)$.
  - Recompute F_{extend}^{tot} with the additional vertex v.
- Return MC_{extend} with the highest F_{extend}^{tot} that is below $\bar{\tau}_{given}$.

5.4 Preliminary Results

With the above implementation, we identified a few interesting genes that drive the similarity and dissimilarity between the Barrett's epithelium and the squamous epithelium. The genes that drive the dissimilarity between the squamous epithelium and the Barrett's epithelium included Human gastrointestinal tumor-associated antigen GA733-1, a marker associated with colon cancer, and Af-17, a putative member of a family of genes involved in cytokinesis and cell cycle control. Genes that drive the similarity between the two tissues included a series of ribosomal subunits (Human acidic ribosomal phosphoprotein P0, Human ribosomal protein L21, Human gene for heterogeneous nuclear ribonucleoprotein core protein A1, Human ribosomal proteins S5, S17, and S29), early stress response genes (Human 90 kD heat shock protein), metabolic enzymes (Human liver mRNA for glyceraldehyde-3-phosphate dehydrogenase), and growth factor responsive genes (Metallopanstimulin 1). The genes that drive the similarity are consistent with the need for cell turnover in these tissues as a result of their constant exposure to acid reflux.

5.5 Directions for Extensions

Our current GENEEXTRACT algorithm is still in its infancy stage. There are many directions for future improvement.

In our current implementation, the total weight $TW(v)$ from each vertex v not in the max-clique MC to vertices in the max-clique are not updated after vertices are added to the extended subgraph MC_{extend}. One possible improvement is to update the total weight to the extended subgraph MC_{extend} every time after a vertex is added.

We assume no ties are present in both Theorem 1 and Theorem 2. The proof of Theorem 1 should go through even if ties are present. However, the proof of Theorem 2 would not go through with the presence of ties. One direction of future work is to modify Theorem 2 to take ties into consideration. However, if we assume that expression levels are real numbers, the no tie assumption is not a significant concern.

In our current implementation, the algorithm starts with a approximate max-clique using the unweighted formulation, and then extends the clique considering the weights on edges. One alternative approach is to directly look for an approximate *weighted* max-clique, which is a complete subgraph with the maximum total weight on its edges (instead of the maximum number of nodes). The current RLS implementation to find max-clique assumes no weights on edges.

There may be many disjoint subsets of genes with high similarity or dissimilarity to each other. Currently, our implementation only returns one such subset. We would like to extend our implementation to rank the disjoint subsets and to return more than one. We can also change the formulation to look for *dense subgraphs* instead of cliques, *i.e.*, $F' = \frac{\sum_{(i,j)\in E'} w_{ij}}{|V'|}$, so that large subsets of genes are automatically preferred in the objective function. Yet another possibility is to search for highly connected subgraphs.

In the formulation of the weighted graph in Theorem 2, the weights of edges do not take into account the degree of concordance or discordance of a pair of genes. Since the Kendall's coefficient of concordance is a *rank* association measure, only the relative ranks matter. For our purpose, we can imagine assigning higher (or lower) weights to pairs of genes that are highly concordant (or discordant).

5.6 Comparison with other approaches

There are many approaches to identify genes that are differentially expressed in two or more types of tissue. Claverie gave many examples of statistical approaches in his review article [Claverie, 1999]. Identifying genes that distinguish two or more tissue types is also known as the *feature extraction* problem in classification. The idea is that the subset of genes that distinguish the two classes (tissue types) should be used as class predictor. For example, [Golub *et al.*, 1999] used the difference of the means in two classes divided by the sum of the standard deviations in the two classes as an estimate for the distinguishing power of a gene.

Our approach is very different: instead of using the distribution of expression levels in each tissue type and in each gene, we compared the expression levels of pairs of genes under the same experiment. Our approach does not assume the expression levels in different experiments to be normalized. In other words, our approach would work even if the overall signal intensities of different chips (experiments) are very different.

In the case of the Barrett's esophagus data, we do not expect the tissue samples to have very different variations. Usually, cancer tissue samples are expected to be more heterogenous. In the Barrett's esophagus data, we have three normal tissue samples, and Barrett's esophagus is premaligant

(not cancer). Moreoever, there are not enough tissue samples (only 3 or 5) from each tissue type to compute robust estimates of the standard deviations. There are only 623 genes (out of 7070 genes) that have high confidence expression levels in all five Barrett's tissue samples and all five squamous tissue samples. The distributions of standard deviations in those 623 genes from the Barrett's and squamous tissue samples are comparable. Therefore, we believe that our GENEEXTRACT algorithm is applicable to the Barrett's esophagus data.

6 Conclusions

In this report, we addressed the three basic questions that motivated this study (see Section 1). We proposed normalization strategies to pre-process the data from two formats of Affymetrix chips, and used the normalized data in our analysis. Pearson's correlation coefficient was used to investigate the similarities of different tissue samples. A novel approach is proposed to filter out genes that are not differentially expressed between different tissue types. Cluster analysis was used to identify tissue specific gene clusters. In addition, a novel algorithm is developed to identify genes that "make Barrett's Barrett's", *i.e.,* genes that make the Barrett's epithelium distinct from (or similar to) each of the other normal gastrointestinal tissues. In terms of future work, we would like to incorporate the extension ideas in Section 5.5. We also believe that our approach to identify genes driving the similarity (or dissimilarity) between different experiments (or tissue types) has many applications. We would also like to explore other applications of our algorithm.

Acknowledgement:
We would like to thank Amir Ben-Dor, Jeremy Buhler, Benno Schwikowski and Rimli Sengupta for their involvement in the preliminary analysis of this work. We would like to thank Richard Karp, Paul Beam and Benno Schwikowski for their advice on practical implementations for the max-clique problem. This work is partially supported by NSF grant DBI-9974498 and NIH grant AG14358. [??? should we also acknowledge other grants or other people???]

References

[Barrett *et al.*, 2000] Barrett, M. T., Yeung, K. Y., Delrow, J., Blount, P. L., Sullivan, R., Zarbl, H., Ruzzo, W. L., Hsu, L., Reid, B. J. and Rabinovitch, P. S. (2000) Transcriptional analysis of barretts epithelium and normal gastrointestinal tissues. Manuscript in preparation.

[Battiti and Protasi, 2000] Battiti, R. and Protasi, M. (2000) Reactive local search for the maximum clique problem. *Algorithmica*.

[Ben-Dor and Yakhini, 1999] Ben-Dor, A. and Yakhini, Z. (1999) Clustering gene expression patterns. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France.

[Blot *et al.*, 1991] Blot, W. J., Devesa, S. S., Kneller, R. W. and Fraumeni, J. F. J. (1991) Rising incidence of adenocarcinoma of the esophagus and gastric cardia. *Journal of the American Medical Association*, **265**, 1287–1289.

[Claverie, 1999] Claverie, J. M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Huamn Molecular Genetics*, **8**, 1821–1832.

[Golub *et al.*, 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**,531–537.

[Hamilton *et al.*, 1988] Hamilton, S. R., Smith, R. R. and Cameron, J. L. (1988) Prevalence and characteristics of barrett esophagus in patients with adenocarcinoma of the esophagus or esophagogastric junction. *Human Pathology*, **19**, 942–948.

[Hastad, 1996] Hastad, J. (1996) Clique is hard to approximate within $n^{1-\epsilon}$. In *Proc. 37th Ann. IEEE Symp. on Foundations of Computer Science*, Vermont, USA.

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

[Johnson and Trick, 1996] Johnson, D. and Trick, M. (eds.) (1996) vol. 26 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, American Mathematical Society, Providence, RI.

[Kendall, 1970] Kendall, M. G. (1970) *Rank Correlation Methods*. London: Griffin.

[Pearson, 1896] Pearson, K. (1896) Mathematical contributions to the theory of evolution, iii. regression, heredity, and panmixia. *Philosophical Transcriptions of the Royal Society*, **A 187**, 253–318.

[Phillips and Wong, 1991] Phillips, R. W. and Wong, R. K. (1991) Barrett's esophagus. natural history, incidence, etiology, and complications. *Gastroenterology Clinics of North America*, **20**, 791–816.

[Silverberg *et al.*, 1990] Silverberg, E., Boring, C. C. and Squires, T. S. (1990) Cancer statistics, 1990. *CA: A Cancer Journal for Clinicians*, **40**, 9–26.

[Snedecor and Cochran, 1980] Snedecor, G. W. and Cochran, W. G. (1980) *Statistical Methods*. Iowa State University Press.

[Yeung *et al.*, 2000] Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2000) Validating clustering for gene expression data. *To appear in Bioinformatics*.

[Zar, 1984] Zar, J. H. (1984) *Biostatistical Analysis*. Prentice Hall.

Appendix

	GAS1	DUO1	BE1	BE2	BE3	BE4	Sq1	Sq2	Sq3	Sq4	GAS2	GAS3	DUO2	DUO3	BE5	Sq5
GAS1	1.000	0.807	0.799	0.848	0.788	0.820	0.753	0.741	0.725	0.764	0.862	0.855	0.762	0.768	0.805	0.741
DUO1	0.807	1.000	0.805	0.814	0.842	0.811	0.736	0.726	0.717	0.737	0.777	0.746	0.864	0.878	0.792	0.719
BE1	0.799	0.805	1.000	0.887	0.826	0.836	0.808	0.810	0.802	0.820	0.766	0.750	0.730	0.748	0.792	0.759
BE2	0.848	0.814	0.887	1.000	0.823	0.883	0.828	0.810	0.821	0.842	0.807	0.789	0.750	0.764	0.825	0.795
BE3	0.788	0.842	0.826	0.823	1.000	0.811	0.744	0.731	0.712	0.757	0.788	0.757	0.763	0.789	0.843	0.722
BE4	0.820	0.811	0.836	0.883	0.811	1.000	0.754	0.753	0.750	0.788	0.789	0.766	0.775	0.787	0.889	0.759
Sq1	0.753	0.736	0.808	0.828	0.744	0.754	1.000	0.902	0.893	0.931	0.732	0.725	0.685	0.704	0.725	0.882
Sq2	0.741	0.726	0.810	0.810	0.731	0.753	0.902	1.000	0.958	0.935	0.697	0.692	0.645	0.664	0.691	0.882
Sq3	0.725	0.717	0.802	0.821	0.712	0.750	0.893	0.958	1.000	0.930	0.696	0.694	0.636	0.660	0.684	0.868
Sq4	0.764	0.737	0.820	0.842	0.757	0.788	0.931	0.935	0.930	1.000	0.732	0.729	0.680	0.699	0.733	0.885
GAS2	0.862	0.777	0.766	0.807	0.788	0.789	0.732	0.697	0.696	0.732	1.000	0.955	0.852	0.861	0.861	0.762
GAS3	0.855	0.746	0.750	0.789	0.757	0.766	0.725	0.692	0.694	0.729	0.955	1.000	0.848	0.854	0.855	0.784
DUO2	0.762	0.864	0.730	0.750	0.763	0.775	0.685	0.645	0.636	0.680	0.852	0.848	1.000	0.967	0.861	0.742
DUO3	0.768	0.878	0.748	0.764	0.789	0.787	0.704	0.664	0.660	0.699	0.861	0.854	0.967	1.000	0.874	0.753
BE5	0.805	0.792	0.792	0.825	0.843	0.889	0.725	0.691	0.684	0.733	0.861	0.855	0.861	0.874	1.000	0.796
Sq5	0.741	0.719	0.759	0.795	0.722	0.759	0.882	0.882	0.868	0.885	0.762	0.784	0.742	0.753	0.796	1.000

Table 2: Correlation coefficients between all 16 experiments (for Section 3).