Hierarchical Cluster Analysis of SAGE Data for Cancer Profiling

Raymond T. Ng
Department of Computer Science
University of British Columbia
rng@cs.ubc.ca

Jörg Sander
Department of Computing Science
University of Alberta
joerg@cs.ualberta.ca

Monica C. Sleumer
Department of Computer Science
University of British Columbia
sleumer@cs.ubc.ca

ABSTRACT

In this paper we present a method for clustering SAGE (Serial Analysis of Gene Expression) data to detect similarities and dissimilarities between different types of cancer on the sub-cellular level. The data, however, is extremely high dimensional, and due to the method of measurement, there are many errors as well as missing values in the data, challenging any clustering algorithm. Therefore, we introduce special pre-processing techniques to reduce these errors and to restore missing data. These techniques are tailored to the process that generates the data, making only very conservative changes. Furthermore, we present a new subspace selection technique to identify a relevant subset of attributes (genes) using the Wilcoxon test. This is a general technique that can be applied to select subspaces for the purpose of clustering whenever some high-level categories of interest are known for the data (such as cancerous and non-cancerous). Finally, we discuss the results of the application of the clustering algorithm OPTICS to the SAGE data, before and after our preprocessing steps.

1. INTRODUCTION

Cancers are typically classified based on macroscopic qualities such as the tissues they developed in. However, the same type of cancer can have different reactions to treatment in different people, and on the other hand, one treatment can be effective for cancers originating in different tissues. A possible explanation for these observations is that cancers, which are similar on a macroscopic level, may in fact be different at the sub-cellular level, and that certain cancers, which originate in different tissues, may actually be similar to each other at the sub-cellular level.

The sub-cellular level of a tissue is characterized by very complex biochemical processes of which only a very small fraction is understood today. Most of the functions of a cell are performed by proteins, which are produced by the cell via a mechanism called *gene expression*.

The general goals of data mining gene expression data are to make discoveries about how the gene expression profiles of diseased tissue are different from those of healthy tissues, how gene expression profiles of diseased tissues are different from each other, and how these differences affect the prognosis of the patient and the success of treatment.

This paper contributes to these goals by proposing a method for clustering gene expression data to detect differences or similarities between normal and cancerous tissues, or between different types of cancerous tissues. In particular, we want to analyze whether different cancerous tissues cluster together, and whether samples of the same cancer type form subclusters.

Some cluster analysis has been done on *microarray* data, but only for single tissue types and involving only a relatively small number of genes. These attempts have been proven to be a promising start, but their results are far from achieving the above goals. The SAGE data is far more complete with respect to the genes and the available data set has been made from many different tissue types, so that it can be used to perform inter-tissue comparisons.

As far as we know, no analysis has yet been performed on the SAGE data set found on the NCBI website. One reason for this is that the data is relatively new and the other reason is that the SAGE data has some unique characteristics with respect to error and dimensionality: each data library (tuple) contains about 10% error due to the method of measurement, and the dimension of the data set is about 350000 which is extremely high. To deal with these properties, which challenge every clustering algorithm, we propose the following preprocessing steps before we apply a clustering algorithm for our analysis:

1. Remove most of the dimensions that contain errors due to sequencing error
2. Impute some of the missing values in truncated libraries
3. Normalize the data
4. Select an appropriate subspace (i.e. a subset of genes) using the Wilcoxon test

The rest of the paper is organized as follows. In section 2, we briefly explain gene expression and how it is measured. Section 3 gives a short overview of related work on the SAGE data and clustering microarray data. We present our cluster analysis of the SAGE data in section 4, where each of the points mentioned above is discussed in detail. Section 5 concludes the paper.

2. GENE EXPRESSION

Gene expression is the process of making mRNA copies of a gene (a part of the DNA strand of the cell) and then translating the mRNA into an amino acid sequence that forms the basis of a protein. Each tissue type in an organism requires different amounts of different proteins to perform its duties. The amount of a particular protein produced by a cell is partly controlled by the number of corresponding mRNA copies. The relative levels of mRNA of each gene in a tissue type are called the tissue's *gene expression profile*. This profile is assumed to be more or less unique for a particular tissue type. Biochemists assume that the vast majority of genes in the human genome are only expressed in one tissue type, and only some "housekeeping genes" are expressed in all cells, such as those that control transcription and translation.

Certain diseases, especially cancer, are caused by a sequence of errors that radically alter the normal pattern of gene expression. There may be a mutation in one or several genes that control the rate of expression of several other genes. For instance, the proteins that prevent uncontrolled cell growth and promote cell death are no longer produced, and the focus of the cell's gene expression machinery switches from producing the proteins appropriate to its tissue type to producing mostly proteins that are needed for cell growth and division. Theoretically, if the gene expression profile of a diseased tissue sample were known, it could be used to find out what has gone wrong in the cell and provide clues as to how to fix it.

There are two major methods of measurement for gene expression data, which produce a snapshot of the gene expression processes in a cell sample, at a certain point in time - the microarray method and the SAGE method.

In the microarray method, the sequences of the mRNAs that are measured must be known in advance. Many single-stranded pieces of DNA that complement these short sequences are printed on a glass chip. The chip is then brought into contact with the mRNAs extracted from a cell sample. The mRNA in the sample binds to its DNA complement on the chip, and causes it to fluoresce. This fluorescence is detected with a laser. The more mRNA of a certain sequence is in the sample, the more its complementary spot on the chip will light up. An advantage of this method is that it is relatively inexpensive to produce large amounts of data. The major disadvantage, however, is that the experimenter must choose the mRNA sequences to be detected in a sample, and the sequences useful for cancer profiling may not be known.

In the SAGE method, all the mRNA of a cell sample is collected and a 10-base tag is extracted from each mRNA (which is a sequence of the 4 bases A, C, G, and T). These tags are linked together in a chain and sequenced. The frequency of each tag is counted, giving the relative levels of the corresponding mRNA in the cell sample. The information gained from performing this technique on a tissue sample is called a SAGE *library*. The advantage of this method is that *all* the mRNA in a tissue sample has a chance of being tagged and sequenced, so the sample is not biased by the experimenter. Furthermore, the sequences of the complete set of mRNAs profiled do not have to be known in advance. This is important, especially for cancer profiling on the gene expression level, because the sequence and purpose of most mRNAs has not been discovered yet. A disadvantage of this technique is that it is very expensive to produce data this way. Therefore, there is not very much publicly available SAGE data yet. When we performed our experiments there were 88 libraries, but new libraries are added every few months. Another disadvantage is that the SAGE data contains more errors than microarray data. Some tags may not be unique, i.e. more than one mRNA may be represented by the same tag, and the method also suffers from problems such as sequencing error and sampling error (for more information on the accuracy of the SAGE method see [16]). However, in spite of the problems, the SAGE method is the only method so far that collects a complete gene expression profile for a cell sample.

There is also a third method to collect gene expression data in which complete segments of mRNA are sequenced instead of just a 10-base tag. However, with this method only a small amount of data can be gathered at very high cost.

3. RELATED WORK

The SAGE method was introduced in [18] in 1995. In [19] the same group proposed that the SAGE method could be used to study the differences between cancerous and normal cells. However, they only provide a brief example of the analysis that could potentially be done but did not provide the results of such an analysis.

In 1999 a website [12] was created as an offshoot of the Cancer Genome Anatomy Project (CGAP) at the National Center for Biotechnology Information (NCBI). CGAP is dedicated to collecting data on the genetics of cancer. An introduction to CGAP and an explanation of its purpose and the tools it contains is presented in [17]. The purpose of the SAGE website is to provide data to the public, so that researchers could benefit from the SAGE technique without having to bear the expense of creating all of the data themselves [10]. Since then, various laboratories have submitted SAGE data to the site, which currently contains a total of 88 different libraries, both cancerous and non-cancerous, from 10 different tissue types. The website also contains various tools to help researchers analyze the data (see [11] for a description). However, all of these tools can only analyze one library or compare two libraries to each other. None of them perform any analysis of the tag frequency distribution on a larger set of libraries.

An excellent summary of the kind of research that could be done with gene expression data can be found in [7]. The authors make several interesting suggestions such as investigating genes which are up-regulated in cancer cells as possible targets for drugs, and finding mRNA sequences whose expression levels vary between cancerous and normal cells and therefore would be good candidates for microarray studies. The paper focuses on complete sequencing of mRNA as the source of the data to be analyzed. The only analysis they actually performed was based on a small amount of such data from various prostate tissue samples taken from 5 patients. On this data set they studied changes in the gene expression of patients over time.

Eisen *et al.* [6] first proposed the use of a clustering algorithm to study gene expression. The authors designed microarrays for the budding yeast genome and a human fibroblast cell line. Their cluster analysis used a hierarchical pairwise average-linkage clustering method, and the correlation coefficient as a distance measure. However, the authors only used their distance function to cluster the genes. The samples were not clustered because they represented an ordered set for which clustering was not meaningful.

Perou *et al.* [14] used the same methodology as Eisen *et al.* to cluster genes using expression profiles of normal breast tissue, breast cancer bulk tissue, and breast cancer cell lines. The authors produced their own microarrays using sequences of genes known or suspected to be involved in cancer. They found some interesting clusters of genes that are co-regulated.

A specifically designed clustering algorithm for clustering genes using gene expression data can be found in [4]. This algorithm is based on a biological model of genes to find natural groups of co-regulated genes, and it is not obvious how to apply the method for finding clusters of tissue samples.

In [2], Alon *et al*. cluster both the genes and the tissue samples for the first time. The authors obtained a number of samples of normal colon bulk tissue, cancerous colon bulk tissue and cancerous colon cell line, and created gene expression profiles with a commercially produced microarray. They also used a hierarchical clustering algorithm and the correlation coefficient as in [6] as a distance function. The clustering of the colon tissue samples was quite successful in separating cancer and normal as well as bulk and cell line tissue.

Golub *et al.* [8] used the same commercially produced microarray as Alon *et al.* to produce gene expression profiles for different types of leukemia. They used self-organizing maps to cluster the data in order to classify these leukemias. In particular, they wanted to find the genetic differences between acute lymphoblastic leukemia and acute myeloid leukemia, based solely on genetic profiles of bone marrow samples, to reduce dependence on expensive laboratory tests. The method was moderately successful in predicting the two leukemia types based on the found clusters. The authors also point out a danger in clustering gene expression data made from solid tumour biopsy samples. Solid tumours are often composed of a mixture of cell types, and are surrounded by various normal non-cancerous cells. Gene expression profiles of tumour biopsy samples could therefore result in clusters being formed on the basis of normal cell contamination.

In [1], Alizadeh *et al*. attempt to find subtypes of a type of lymphoma that are different from the currently recognized subtypes, since the current subtypes are not helpful in determining a patient's prognosis or best course of treatment. They designed and produced their own microarray, made with sequences that are known or suspected to play a role in lymphoma. This microarray was then used to create gene expression profiles of both normal and cancerous lymphocyte samples, which were analyzed with the same clustering method as in Eisen *et al.* The clusters they discovered were actually slightly better at predicting the eventual prognosis of patients than the International Prognostic Index. The authors conclude that only on a larger data set could the method reveal more subtypes of lymphoma and increase the prediction accuracy for prognoses of patients.

4. CLUSTER ANALYSIS OF SAGE DATA

The following analysis is based on 88 SAGE libraries which were publicly available on the NCBI SAGE website. Each of these SAGE libraries is made of a sample from one of the following tissues: brain, breast, prostate, ovary, colon, pancreas, vascular, skin, or kidney. The only other information consistently included with each library is whether it was made from cancerous or normal tissue, and whether it was made from bulk tissue *in vivo* or a cell line grown *in vitro*. The data in each library includes a list of 10-base tags that were detected in the sample, and the number of times each tag was detected. The number of *unique tags* in a library is the number of different tags that were detected in the sample, each of which was detected with some integer frequency. The *total number of tags* is the sum of all the frequencies of all the tags in a library.

Analyzing the process by which the SAGE data is produced, it is clear that the data has four major problems. First, the SAGE method is highly prone to sequencing error, which creates a large amount of noise and obscures the clustering structure. Second, the libraries differ largely in terms of total number of tags due to the degree of completeness of the sequencing. Third, some of the SAGE libraries are extremely short, having 0 values for most of the tags. This is also due to partial sequencing, which is extremely incomplete in some cases. Fourth, the data is extremely high dimensional, and it is not clear which dimensions are relevant for cancer profiling. Therefore, we designed four pre-processing steps to deal with these problems: error removal, normalization, missing tag imputation, and subspace selection. Each of these steps is discussed in detail below.

There are several different possibilities for a distance measure between SAGE libraries, such as Manhattan distance, Euclidean distance, the angle, and the correlation coefficient. The one that is most commonly used in the literature for microarray data is the correlation coefficient. We clustered the SAGE libraries using all of the mentioned distance functions and found that the clustering structures are typically very robust with respect to the distance function. However, the correlation coefficient always resulted in the most pronounced structure, and is also consistent with previously published work. Therefore, we will present our results using the correlation coefficient.

4.1 Error Removal and Normalization

4.1.1 Error Removal

Because 10% of the total number of tags in each SAGE library are expected to contain sequencing errors, and since these errors result in noise and increase the dimensionality of the data, our first pre-processing step is error removal. Within one library, up to 80% of the unique tags have a frequency of 1, but these make up about 20% of the total number of tags. Therefore, removing all tags of frequency 1 from every library would remove too many tags. Some of the single frequency tags represent genuine low frequency mRNAs. However, it is impossible to tell which of the single frequency tags are errors by only looking at *one* library.

Since the size of the set of possible tags is $4^{10} \approx 10^6$, and only about 3×10^5 of these tags have ever appeared in any library, a sequence error in a tag has most likely resulted in the creation of a completely new tag that never appears in any other library. Therefore, our method of removing erroneous tags without removing legitimate single-frequency tags is as follows: only remove those tags that have a frequency of no more than 1 in any of the libraries. This way, tags that have a frequency of 1 in some libraries but a higher frequency in other libraries are not removed.

The size of the union of all the unique tags (which is the dimensionality of the subsequent space upon which the libraries are clustered) detected in all the libraries before this removal is about 350000; after the removal it is 58524. Therefore, this method reduces the dimensionality of the data set as well as the noise, and removes between 5% and 15% of the total number of tags in each library.

4.1.2 Normalization

The 88 SAGE libraries upon which we base our analysis were made by different institutions. Because the amount of resources spent by these institutions for sequencing varies widely, each library has a different total number of tags. As a consequence, it is not meaningful to use some of the distance functions such as the Euclidean distance directly to compare the libraries to each other, since their tag frequencies are not on the same scale. To be able to use these distance functions, we normalized the tag frequencies by simply scaling them all up to the same total number of tags (and normalization is also necessary for the subspace selection technique presented below). The value we scaled them up to is 300000, which is the estimated number of mRNAs per cell [19]. For other distance functions such as the correlation coefficient this type of scaling does not have any effect.

4.1.3 Preliminary Analysis Using CLARANS and OPTICS

In our early attempts to cluster the SAGE libraries, we decided to use CLARANS [13], a *k*-medoids algorithm, and we focused on a single tissue type, i.e. 14 breast tissue libraries that were available at the time. We applied some simple screening procedures to reduce the dimensionality of the data set, such as requiring a minimum frequency threshold for tags and considering only the top *n* most frequent tags from each library, where *n* was between 30 and 100. The error removal step described above is obviously not necessary for this type of screening since it automatically excludes very low frequency tags, however, the error removal step was applied for all subsequent analyses where other tag selection methods were used.

We tested several distance functions such as Euclidean distance, correlation coefficient, angle, and degree of overlap between the top *n* most frequent tags in the compared libraries. This analysis showed that the breast tissue libraries form clusters, but these clusters were formed according to tissue source (cell line or bulk tissue) and not according to neoplastic state (cancer or normal). All of the bulk tissue-derived libraries went into one cluster while all the cell line-derived libraries went into another cluster. This showed that there are differences between bulk tissue and cell line libraries which should not be ignored. But even after eliminating all tags which were only found in bulk tissues, the success in separating cancerous breast tissue from normal breast tissue was only moderate. We also found that CLARANS was not effective in clustering all 88 libraries at once, regardless of the distance measure used. CLARANS always creates the user-determined number of clusters and gives no hints as to what an appropriate number of clusters would be. We tried several different numbers of clusters *k*, but the clustering structure was extremely weak in all cases.

Because of these weak results, and because we actually wanted to see whether some types of cancers were related to each other at the gene expression level as well as find clusters within one category, we decided to use a hierarchical clustering algorithm for a more detailed analysis. Hierarchical algorithms were also used in previously published papers on the clustering of microarray data. Our analysis is based on the hierarchical clustering algorithm OPTICS [3], which improves over traditional hierarchical clustering algorithms in several ways:

- Although similar to the *single-link* or *nearest-neighbor* method, a parameter of OPTICS can be used to weaken the single-link effect. (This parameter can also be set in way that the result of OPTICS is equivalent to the result of the single-link method.)
- The output of the OPTICS algorithm is a *reachability plot*, which gives a clearer view of the hierarchical clustering structure and the density of the clusters than the typical dendrogram produced by traditional hierarchical clustering methods. (However, a reachability plot can be converted to a dendrogram if necessary).

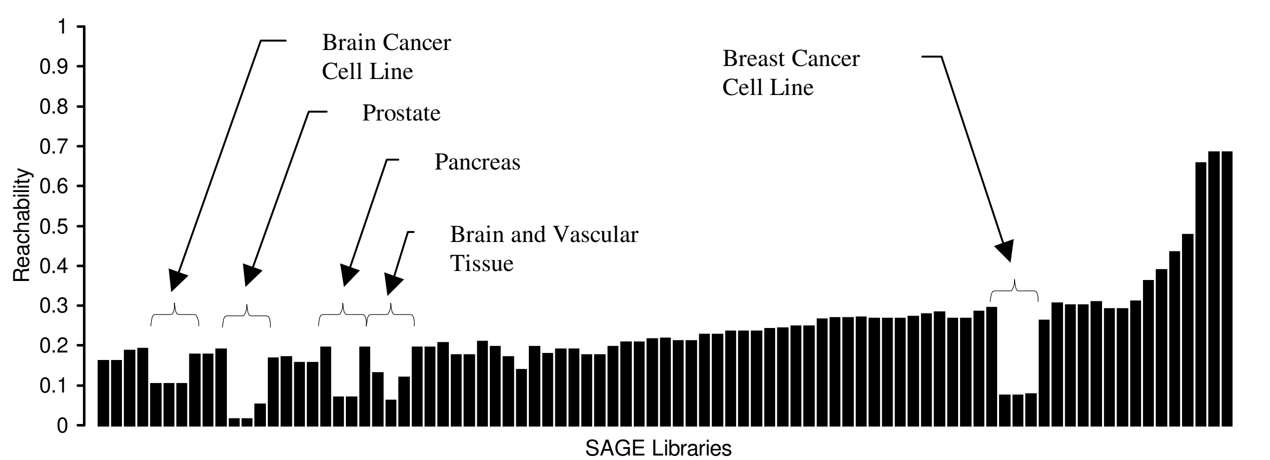

Figure 1: Result of OPTICS on the SAGE data after error removal and normalization.

- OPTICS is more efficient on large data sets than traditional hierarchical clustering algorithms, if the dimension of the data set is not too high. (For the SAGE data set, however, this aspect is not relevant since the data has very few tuples and is ultra-high dimensional.)

The reachability plot is a simple bar plot which visualizes a *cluster ordering* of the data set. It shows simultaneously which clusters are formed by which points (easily recognized as "valleys" in the plot), how dense the clusters are in relation to each other (the deeper the "valley", the more dense the cluster), a lower bound on the distance between two clusters (the tallest bar between two "valleys"), the hierarchical structure of the clusters ("nested valleys"), and which points are outliers (very large bars in the plot, not at the border of a cluster). Applying OPTICS naively to the SAGE data, however, also does not produce a strong clustering structure. Figure 1 shows the result using the correlation coefficient as a distance function, which produced the most pronounced clustering result compared to all other distance functions we applied. There are very few clusters: a brain cancer cell line cluster, a prostate cluster, a pancreas cluster, a mixed cluster containing brain and vascular tissue, and a breast cancer cell line cluster. Each of these clusters contains only 3 to 5 libraries, all of which are of a single type of tissue (with the exception of the cluster containing both brain and vascular tissue). The other 56 libraries do not group into pronounced clusters. This weak result is consistent with our preliminary results using CLARANS, and indicates that we have to address the problem of missing values and suitable subspace selection. We can observe again that several tissue types tend to separate by source rather than by neoplastic state. For instance, the brain cancer cell line libraries do not cluster together with the brain cancer bulk tissue, which shows that the bulk tissue samples contain considerable contamination.

4.2 Missing Tag Imputation

After the error removal outlined in 4.1.1, the total number of tags in the libraries varies from 1293 to 79498. The libraries with a very low number of tags can be thought of as "truncated" - there weren't enough tags sequenced in these libraries to be able to detect some of the low-frequency mRNAs. Therefore, we used a method to impute missing values which is similar to techniques for filling in missing answers on incomplete surveys [15]. Libraries that have missing values can be adjusted using other libraries in the same category, i.e. from the same tissue, source, and neoplastic state. Otherwise, these libraries will never be able to cluster properly and the information contained in them would be lost. The missing tags were replaced as follows:

- Identify libraries with a low total number of tags
- For each of these,
  - Create a list of libraries in the same category, which have a sufficiently high number of tags.
  - Scale them down so that the total number of tags matches the total number of tags of the low tag library.
  - Replace each frequency 0 in the low-tag library with the lowest non-zero frequency of that tag in the scaled down libraries.

Our intention was to solve the problem of the truncated libraries as conservatively as possible, which is why we replaced the zeros with the lowest possible value from the scaled down libraries instead of with the average.

4.2.1 Result of OPTICS After Tag Imputation

After the low frequency tags were removed, the missing values were filled in and the libraries were normalized, the resulting data was clustered by OPTICS. Figure 2 shows the result, again using the correlation coefficient as a distance measure. Basically, the same clustering structure as before is found, only the pancreas cluster now includes 2 additional libraries. This result shows that our imputation of missing values did not change the data radically with respect to the clustering structure and did not create any "artificial" clusters by making libraries too similar to each other. The imputation of missing values has, however, important benefits for subspace selection, as we will see in the next section.

4.3 Subspace Selection

To focus our analysis on those tags which identify cancer types, we determined that there were two major types of tags that

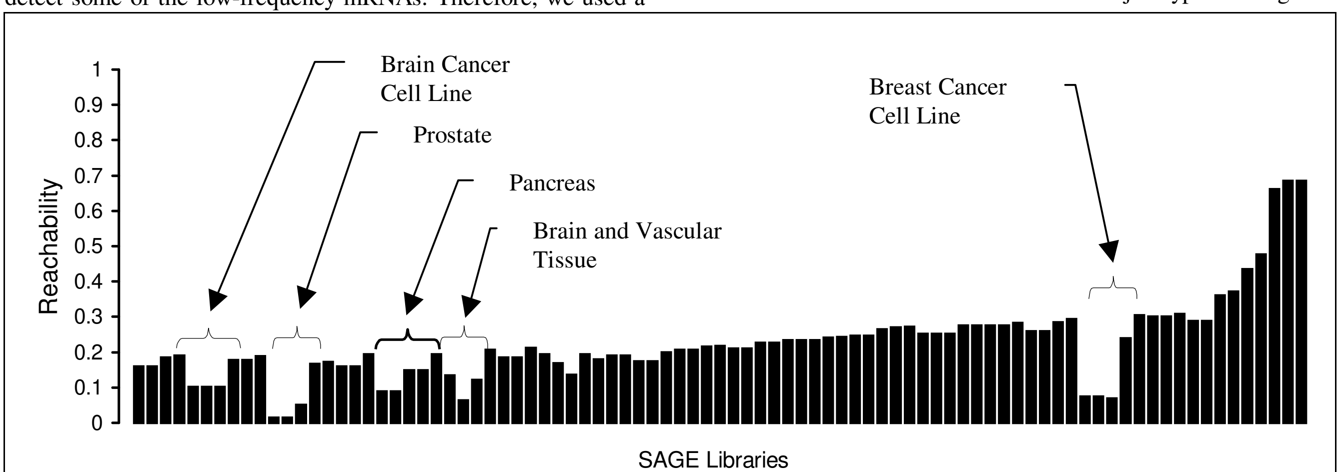

Figure 2: Result of OPTICS on the SAGE data after error removal, normalization, and missing tag imputation.

needed to be removed from all the libraries:

- tags which have similar expression levels in all or most of the libraries
- tags which come from contaminating cells in bulk tissue samples

We addressed these issues by using the Wilcoxon test to determine which tags should be removed from the data in order to improve the clustering structure.

The Wilcoxon test (see e.g. [5]) is used to test whether two samples are taken from the same population, i.e., the null hypothesis H_0: $F_X(x) = F_Y(y)$ is tested on the basis of one sample $(x_1, \ldots, x_{n1})$ from X and one sample $(y_1, \ldots, y_{n2})$ from Y. The values of the two samples are taken together and sorted in descending order, resulting in a joint sequence of values from both X and Y. Each pair (x_i, y_j), $x_i \in X$, $y_j \in Y$, with $y_j < x_i$ in the resulting sequence is called an *inversion*, and the expected number of inversions assuming the hypothesis H_0 is $n_1n_2/2$. Thus, this hypothesis can be rejected if the number of actually occurring inversions in the joint sequence differs too much from its expectation. Two important properties of the Wilcoxon test in our context are that it is a non-parametric statistical test, i.e., it makes no assumption concerning the distribution of the two populations, and that it can be successfully applied even to very small samples.

To select a relevant subspace of tags for our application, we first applied the test to all tissue types to determine the tags which have significantly different expression levels in cancerous versus normal cells (using a significance level $\alpha = 0.01$). We then used the test again to determine which tags have significantly different expression levels in bulk tissue versus cell line and subtracted these tags from the set of tags selected in the first step.

4.3.1 *Result of OPTICS on the Selected Subspace*

First, we applied the Wilcoxon test to the SAGE data after error removal and normalization, but without imputing missing values. The Wilcoxon test selected 45354 tags out of 58524 when testing cancer versus normal. 722 tags out of the original 58524, were selected when testing bulk tissue versus cell line. Only 49 of these tags were also present in the 45354 tags selected by the first test. We removed these tags from our dataset, leaving us with a dimension of 45305, and clustered the data set with OPTICS. The result is shown in Figure 3. This subspace selection resulted in a more pronounced clustering structure than the previous experiments. For instance, the figure shows a large prostate tissue cluster containing 13 libraries. However, most of the clusters are still very small and several of them are mixed with respect to tissue type and neoplastic state. Furthermore, a large number of libraries are not contained in any significant cluster.

The result of OPTICS on the SAGE data, using all pre-processing steps in combination, is shown in Figure 4. In this case, the Wilcoxon test selected 40016 tags when testing cancer versus normal and 186 of these were also selected when testing bulk tissue versus cell line (i.e. 39830 tags remained). In this result the clusters are much larger and all of them except for one are pure in terms of tissue type and neoplastic state.

Eight dense clusters formed: ovarian cancer cell line, brain cancer cell line, brain cancer bulk tissue, prostate tissue, pancreatic cancer, breast cancer cell line, normal brain, and normal breast bulk tissue. The ovarian cancer cluster did not appear in the any of the previous results, and is surprisingly close to the brain cancer cell line cluster, which suggests that these two cancers may be related in some way. It would be of further interest to investigate why these two types of cancers clustered together and which tags they are similar for. We also see that we have succeeded to some degree in focusing our data on tags that are related to cancer growth and development. For instance, brain cancer bulk tissue and normal brain bulk tissue have been separated. But some further analysis may be needed since the brain cancer bulk tissue and brain cancer cell line clusters are in two different clusters, which means that there are still significant differences between bulk and cell line tissue in the selected subspace. The prostate tissue cluster is almost pure - only 3 out of 12 libraries are normal while the rest are prostate cancer. However, since all of them were made from bulk tissue samples, it is likely that these 3 normal libraries contain a lot of contamination. We also see a larger breast cancer cell line cluster than in any of our previous results, and a normal brain cluster and a pancreatic cancer cluster have appeared for the first time.

This strong result is only possible when all four preprocessing steps – error removal, imputation of missing values,

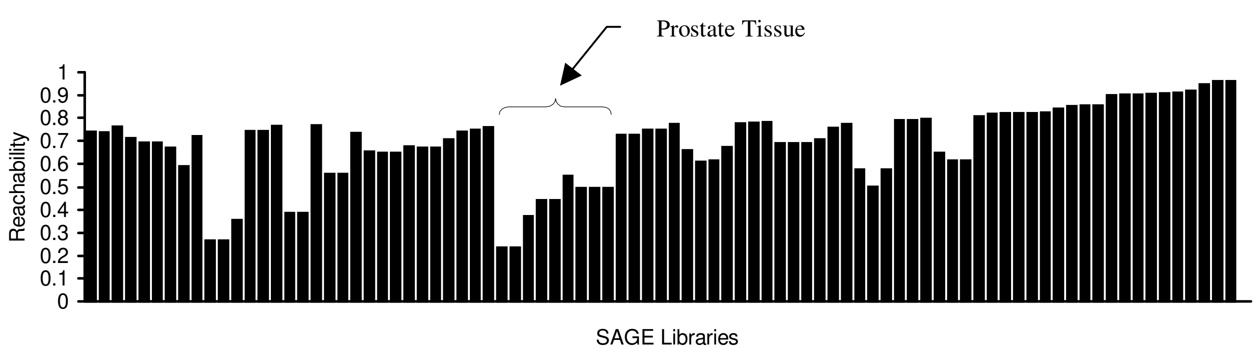

Figure 3: Result of OPTICS on the SAGE on the subspace selected by the Wilcoxon test (including error removal, normalization, but NO missing tag imputation)

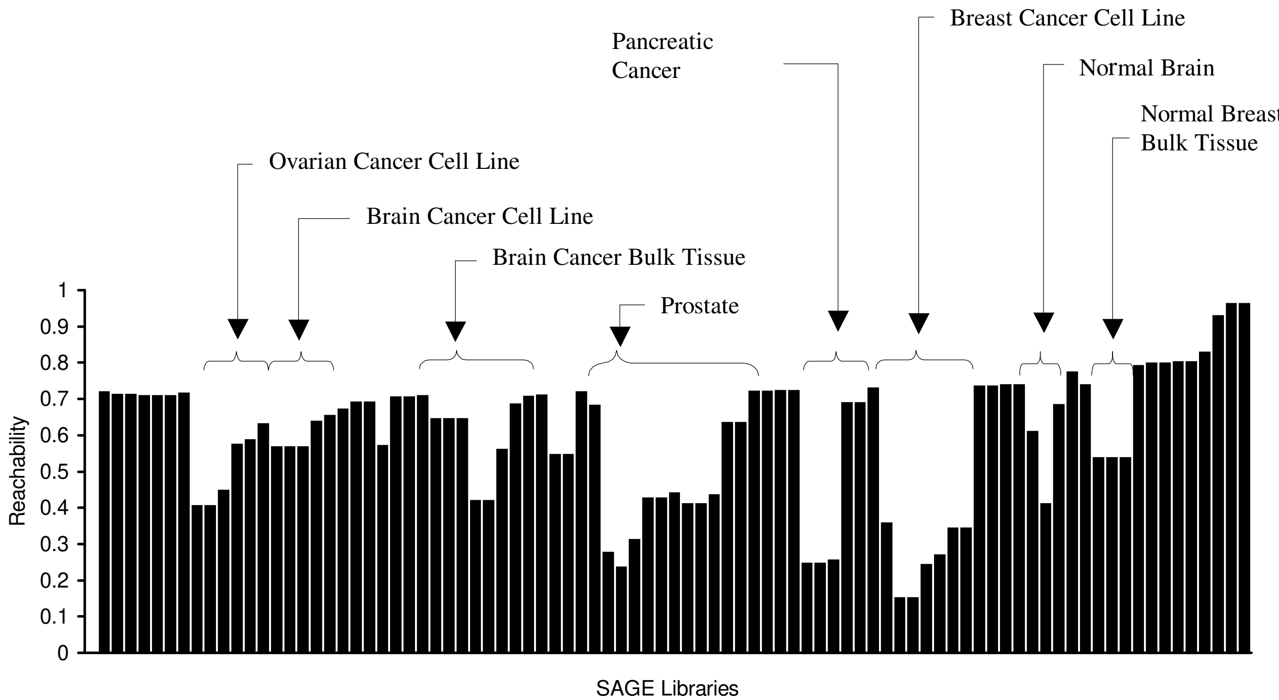

Figure 4: Result of OPTICS on the SAGE on the subspace selected by the Wilcoxon test (including error removal, normalization, and missing tag imputation)

normalization, and subspace selection by the Wilcoxon test – are combined. However, we did not find any sign of a type of cancer resolving into subtypes. This is probably due to the fact that there are not enough libraries of any one type to detect subtypes if they exist. Only when more SAGE libraries for cancerous samples of the different tissue types are available, will it be possible to investigate this problem more thoroughly.

5. CONCLUSIONS

In this paper we proposed a method for clustering SAGE data to detect similarities and dissimilarities between different types of cancer on the sub-cellular level. We introduced four pre-processing steps to reduce errors, restore missing data, normalize the data, and select an appropriate subspace based on the Wilcoxon test. After all of these steps are performed, our clustering algorithm OPTICS produced a promising hierarchical clustering structure in which the SAGE libraries are grouped according to tissue type and neoplastic state, showing a possible relationship between brain and ovarian cancer.

This is, however, preliminary work in several respects. Further research is necessary to investigate the significance of our findings, and more data is required before subtypes of cancers can be discovered. We based our analysis on 88 SAGE libraries, but more libraries are added every few months. However, SAGE data is expensive to produce and it may take some time before a large enough data set is available. Nevertheless, producing more SAGE data is important because, as our experiments indicate, it contains valuable information that is not available elsewhere.

This type of analysis may eventually allow researchers to address important questions such as: whether SAGE results are highly reproducible, whether outliers in the clustering result represent errors or genuine genetic variations, and what the similarities and differences between clusters represent on the biochemical level.

The proposed subspace selection technique is not restricted to the SAGE data. It can also be applied to other high dimensional data sets for the purpose of clustering, whenever some high-level categories of interest are known for the data (such as cancerous and non-cancerous).

There are several possibilities for future research. So far we have applied our method to the SAGE data and have found a number of tags that appear to be significant for cancer profiling. These results could be used to produce a microarray chip containing the corresponding genes for producing a higher quality data set. We could also apply our method to existing microarray data. As another possible technique to select meaningful subspaces in the presence of categories we intend to investigate FASCICLES [9]. The FASCICLES algorithm can find a set of tuples that are similar as well as identify which dimensions (attributes) they are similar for. For instance, this could be utilized to determine more tags that distinguish between bulk tissue and cell line to reduce the influence of contaminations in bulk tissue. It would also be interesting to consider the mapping of tags into references to genes using the tag-to-gene mapper to get information about the biochemical pathways. Taking into account known relationships between genes, similarities between tags could be defined for a more sophisticated distance function.

6. ACKNOWLEDGMENT

We would like to thank Steve Jones and his staff at the BC Genome Sequence Centre for providing us with invaluable information and advice.

7. REFERENCES

[1] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Welsenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., Staudt, L. M.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, Vol. 403, 3 February 2000, pp. 503-511.

[2] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci USA*, Vol 96, June 1999, pp. 6745-6750.

[3] Ankerst, M., Breunig, M., Kriegel, H.-P., Sander, J. 1999.: OPTICS: Ordering Points to Identify the Clustering Structure. *Proc. 1999 SIGMOD*, ACM Press, New York, NY, pp. 49-60.

[4] Ben-Dor, A., Shamir, R., Yahkini, Z.: Clustering gene expression patterns. *J. Comput. Biol.*, Vol. 6, 1999, pp. 281-297.

[5] Bronstein, I. N. and Semendyayev, K. A.: *Handbook of Mathematics*, Springer 1998.

[6] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci USA*, Vol. 95, December 1998, pp. 14 863-14 868.

[7] Emmert-Buck, M. R., Strausberg, R. L., Krizman, D. B., Bonaldo, M. F., Bonner, R. F., Bostwick, D. G., Brown, M. R., Buetow, K. H., Chuaqui, R. F., Cole, K. A., Duray, P. H., Englert, C. R., Gillespie, J. W., Greenhut, S., Grouse, L., Hillier, L. W., Katz, K. S., Klausner, R. D., Kuznetzov, V., Lash, A. E., Lennon, G., Linehan, W. M., Liotta, L. A., Marra, M. A., Munson, P. J., Ornstein, D. K., Prabhu, V. V., Prange, C., Schuler, G. D., Soares, M. B., Tolstoshev, C. M., Vocke, C. D., Waterston, R. H.: Molecular Profiling of Clinical Tissue Specimens: Feasibility and Applications. *American Journal of Pathology*, Vol. 156, No. 4, April 2000, pp. 1109-1115.

[8] 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., Lander, E. S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, Vol. 286, 15 October 1999, pp. 531-537.

[9] Jagadish, H. V., Madar, J., Ng, R. T.: Semantic Compression and Pattern Extraction with Fascicles. *Proc. VLDB*, Morgan Kaufmann Publishers, San Francisco, CA, 1999, pp. 186-197.

[10] Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., Riggins, G. J.: A Public Database for Gene Expression in Human Cancers. *Cancer Research*, Vol. 59, November 1, 1999, pp. 5403-5407.

[11] Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., Altschul, S. F.: SAGEmap: A Public Gene Expression Resource. *Genome Research*, Vol. 10, No. 7, 2000, pp. 1051-1060.

[12] NCBI (National Center for Biotechnology Information): *SAGE: Measuring Gene Expression.* http://www.ncbi.nlm.nih.gov/SAGE

[13] Ng, R.T. and Han, J.: Efficient and Effective Clustering Methods for Spatial Data Mining. *Proc. VLDB*, Santiago, Chile, Morgan Kaufmann Publishers, San Francisco, CA, 1994, pp.144-155.

[14] Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., Botstein, D.: Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci USA*, Vol 96, August 1999, pp. 9212-9217.

[15] Statistical Services, University of Texas: *General FAQ #25: Handling missing or incomplete data*, http://www.utexas.edu/cc/faqs/stat/general/gen25.html

[16] Stollberg, J., Urschitz, J., Urban, Z., Boyd, C. D.: A Quantitative Evaluation of SAGE. *Genome Research*, Vol. 10, 2000, pp. 1241-1248.

[17] Strausberg, R. L, Buetow, K. H., Emmert-Buck, M. R., Klausner, R. D.: The Cancer Genome Anatomy Project: Building an Annotated Index. *Trends in Genetics*, Vol. 16, No. 3, March 2000, pp. 103-106.

[18] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K.W.: Serial analysis of gene expression. *Science*, Vol. 270, Oct 20 1995, pp 484-487.

[19] Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., Kinzler, K. W.: Gene Expression Profiles in Normal and Cancer Cells. *Science*, Vol. 276, 23 May 1997, pp. 1268-1272.