Application of Network Clustering Methods to Gene Expression Profiling Data

Rosemary Braun
National Cancer Institute, National Institutes of Health, Bethesda, MD

Eric Kasper
University of Minnesota

Elliot Martin
Complexity Science Group, Department of Physics and Astronomy,
University of Calgary, Calgary, Alberta, Canada, T2N 1N4

Corinne Teeter
University of California, San Diego and
Salk Institute for Biological Studies
(Dated: November 22, 2009)

We present the application of two new unsupervised clustering methods to gene expression data. The methods—partition decoupling (PDM) and commute time distance (CTD)—are based upon the spectral analysis of the correlation structure of the data. Because they have the ability to reveal non-linear and non-convex geometries present in the data, spectral clustering based approaches are an improvement over typical linear clustering algorithms. Here, we apply these methods to a publicly-available gene expression data set, and demonstrate that we are able to identify cell types and treatments with higher accuracy than is obtained through other approaches.

I. INTRODUCTION

Gene expression profiling is an experimental technique that uses RNA microarrays (small array-like chips functionalized with short RNA fragments) to assay the abundance of each gene’s expression in a given sample. Since their first use nearly fifteen years ago [1], microarray gene profiling experiments have become a ubiquitous tool in the study of disease. The vast number of gene transcripts assayed by modern microarrays (10^5–10^6) has driven forward our understanding of biological processes tremendously, both by elucidating mechanisms at play in specific phenotypes and by revealing previously unknown regulatory mechanisms at play in all cells. However, this high-dimensional data—often with many more variables than samples, and subject to noise—presents analytical challenges.
In the most common analyses of microarray data, each gene is tested individually for association with the phenotype of interest, adjusting at the end for the vast number of multiple comparisons. Follow-up analyses, such as gene-set enrichment analysis [2], are carried out once univariate association with phenotype has been measured. While these gene-at-a-time approaches have been fruitful, they also have the potential to miss mechanisms which can be affected by a change in any one of several genes (such that no single alteration reaches significance) as well as mechanisms which require the concerted activity of multiple genes to produce a specific phenotype. Building on the hypothesis that functionally related genes will display correlated gene expression patterns, clustering analysis has emerged as a tool in gene expression profiling. Most of the clustering approaches implemented today are distance-based, including as hierarchical clustering [3], k-means clustering [4, 5] and Self Organizing Maps [6]; a brief overview may be found in [7]. Of these, k-means appears to perform the best [7, 8], and these methods are simple, visually appealing, and have identified a number of co-regulated genes and phenotype classes. However, they have the drawback of both being sensitive to the extensive noise present in microarray data and being unable to distinguish nonlinear or nonconvex geometries in the feature space.

Spectral clustering approaches offer several advantages over previously used algorithms. They utilize a low-dimensional embedding of the data (thus excluding the noisy, high-frequency components) and can detect geometries which other methods cannot. In addition, they are capable of exploring multiple geometrical scales within data. This attribute may be particularly important in the analysis of gene data where biological relationships between genes do not necessarily exist on the same scale. In this paper we apply two promising spectral clustering methods to gene-expression data: the partition decoupling method (PDM) and the commute time distance (CTD) method.

PDM is a statistical learning technique that reveals the scale-dependent geometry of a network structure using hierarchical spectral clustering. The method was described in detail in [9], where it was applied to the market behavior of the New York Stock Exchange revealing the dynamics of stock prices as a linear combination of effects at various scales (including market, sector and industry). Subsequently, it has been applied [10] to voting data from historical roll call votes from the United States Congress, providing a geometric model for parliamentary voting that articulates the complex structure of political ideology.

CTD is a relatively new and novel method that provides another measure of distance. The commute time between two points i and j, \(\eta(i, j)\), is the expected number of steps a random walker would take to leave i, reach j at least once, and return to i. This means that we can reduce the commute time between i and j by either increasing the strength of the correlation between them, or by adding more paths between them. As in spectral clustering, CTD uses a low dimensional embedding utilizing eigenvalues of the commute time matrix associated with the lowest eigenvalues. Similar attempts at using this kind of technique to cluster data have been applied to collaborative recommendations of movies [11], and to finding related Wikipedia pages [12].

Because cancers are diseases with neither a single etiologic origin nor clear determinant of
outcome, substantial literature is devoted to gene expression profiling of tumor cells. The molecular heterogeneity of cancer is the result of the acquisition of multiple genetic alterations that contribute to the development of the tumor, and the molecular heterogeneity is reflected in the diversity of cancer phenotypes. We might thus expect a complex geometry of relationships between genes and between samples, suggesting that PDM, and CTD may be a suitable tools in the analysis of cancer gene expression data. In this paper, we apply PDM and CTD to gene expression data, revealing phenotype-related structures which were undetectable using other clustering methods.

II. METHODS

This paper will address the application of two spectral clustering methods to gene array data; the partition decoupling method (PDM) and the commute time distance (CTD) method. These methods seem particularly promising in their application to microarray data due to their ability to yield multiscale topological descriptions of network structure and thus uncover relationships in gene expressions which would otherwise go unnoticed. Both approaches utilize the spectral decomposition of the Laplacian matrix; however, the two techniques differ in their methods for finding structure within data. PDM iteratively utilizes two submethods to first find the dominant structure within the system, and then removes this structure to reveal finer-scale relationships within the data. CTD aims to find structure by randomly walking different paths through a weighted and undirected graph in order to find the “lowest energy” path through the network, under the assumption that this path is representative of relationships amongst gene expression profiles.

A. Partition Decoupling Method (PDM)

The partition decoupling method consists, generally speaking, of two iterated steps: the spectral clustering of the data, followed by the removal of the cluster centroids (“scrubbing”) such that another clustering step may be performed on the residuals. The two steps are repeated until the residuals are indistinguishable from noise. By performing successive clustering steps, factors contributing to the partitioning of the data at different scales may be revealed.

1. Spectral Clustering

The first step, spectral clustering, serves to identify clusters of samples in high-dimensional space, thereby permitting dimension reduction from the number of features (e.g., genes) to the number of clusters. The motivation is simple: given a set of data points and some measure of similarity \( s_{i,j} \) between each pair, we wish to partition the data such that points within one cluster are similar to each other. Because the data may readily be represented as a complete graph in which nodes correspond to samples and edge weights correspond to the similarity between two nodes, we
can reformulate this problem as the min-cut problem: cutting the graph across edges with low weights, so as to generate several subgraphs for which the similarity between nodes is high. It has been demonstrated [13, 14] that the min-cut problem may be solved through eigendecomposition of the graph Laplacian, where the Laplacian is a matrix formed from the similarity matrix $S$ (comprised of $s_{i,j}$) and the diagonal degree matrix $D$ with elements $d_i = \sum_j s_{i,j}$. The Laplacian may be defined several ways, corresponding to different normalizations; in this work, we use the symmetric normalization, i.e.,

$$L = I - D^{-1/2} S D^{-1/2}. \tag{1}$$

The similarity measure between two data points is, in this work, computed from their correlation $\rho_{i,j}$ by first converting the correlation to a chord distance on the unit sphere and then exponentiating,

$$s_{i,j} = \exp \left( -\left( \frac{\sin \left( \arccos (\rho_{i,j}) / 2 \right)}{\sigma^2} \right)^2 \right), \tag{2}$$

where the scaling parameter $\sigma$ may be tuned to reveal structure at various scales of the system.

The spectrum of the graph, given by eigendecomposition of $L$, contains information regarding the graph connectivity. Specifically, the number of zero-value eigenvalues corresponds to the number of connected components; since we have a complete graph, there will be exactly one. The second-smallest eigenvalue and its associated eigenvector (the so-called Fiedler value $\lambda_1$ and vector $v_1$) encodes the coarse geometry of the data, with successive eigenvectors describing finer (higher-frequency) structure. By embedding the data into a smaller-dimensional space defined by the low-frequency eigenvectors and clustering the embedded data, the geometry of the data may be revealed.

The embedded data may then be clustered using a clustering algorithm such as $k$-means [4]. Because $k$-means clustering is by nature stochastic [4], multiple $k$-means runs are performed and the clustering yielding the smallest within-cluster sum of squares is chosen. In order to use $k$ means on the embedded data, two parameters need to be chosen: the number of eigenvectors $l$ to use (that is, the dimensionality of the embedded data) and the number of clusters $k$ into which the data will be clustered.

**a. Optimization of $l$.** The optimal dimensionality of the embedded data is obtained by comparing the eigenvalues of the Laplacian to the distribution of Fiedler values obtained from null data. The motivation of this approach follows from the observation that the size of eigenvalues corresponds to the degree of structure (see [14]), with smaller eigenvalues corresponding to greater structure. Specifically, we wish to construct a distribution of Fiedler values – eigenvalues encoding the coarsest geometry of the system – and select the eigenvalues from the true data that are significantly small with respect to this distribution. In doing so, we select the eigenvalues that indicate greater structure than would be expected by chance alone.

Here, the null distribution of Fiedler values was obtained through row-resampling of the data, followed by recomputation of $L$. By resampling the data within rows, correlation between the
samples (columns) is destroyed while preserving the marginal distributions of the rows (gene expressions). By performing this resampling multiple times and recomputing the null Fiedler value $\lambda_1^\prime$, a distribution is obtained; eigenvalues $\lambda$ falling below the 0.05 quantile of the distribution of $\lambda_1^\prime$ are denoted as significant. In the case where no non-zero $\lambda$ meets this threshold, we conclude that there is no structure present in our data which is distinguishable from noise, and halt.

Although row-wise resampling preserves the distribution within each feature, it has two potential drawbacks: first, row-wise resampling is computationally expensive, and second, it destroys correlations between rows (genes), not only between columns (samples). An alternative approach involves resampling the similarities $s_{i,j}$ themselves, which may be thought of as “rewiring” the network. Because this approach retains greater similarity to the original data, the significance threshold obtained is smaller; however, in the case of the data described here, the clustering results were identical, suggesting that the relevant structure is contained within the smaller set of eigenvectors. While a full comparison of these approaches is not presented in this paper, it should be noted that the method for constructing the null distribution of Fiedler values must be carefully chosen with respect to the underlying process generating the data.

b. Optimization of $k$. Methods for obtaining the number of clusters $k$ suitable for partitioning a data set are an open research question (see, e.g., [14, 15] and references therein). We briefly describe several approaches here, noting once again that there is no one-size-fits-all solution.

The first approach exploits the property [10, 14] that clustering the entries in the Fiedler vector yields the best decomposition of the network components. Consequently, one can use the number peaks in the density of the Fiedler vector—that is, the number of values about which the elements of $v_1$ are clustered—as the number of clusters. To obtain this value, we fit a Gaussian mixture model [16] with 2–30 components (assuming unequal variances), computing the Bayesian Information Criterion (BIC) for each mixture model and choosing the optimum (for details of the implementation, see [17, 18]).

Another approach attempts to optimize the robustness of the clustering with respect to increasing $k$. The underlying idea here is that, for the correct number of clusters, incrementing to $k + 1$ clusters will leave the majority of the samples in the $k$ clusters as previously obtained, assigning only outliers to the new cluster. In practice, this is done by computing the clustering for $k = 2$–30 clusters, keeping track for each $k$ of how many additional clusters could be added before one of the original clusters split such that no new cluster has >50% of the data of the original cluster. The $k$ for which this statistic is at a maximum is then considered optimal.

A related approach takes advantage of the expectation that, at the correct value of $k$, the clustering will be stable even with a subset of the data, whereas if $k$ is too large, the clustering will fluctuate strongly with small differences in the data. In this approach, one splits the rows of the data matrix at random into two subsets and clusters the columns for each; the clusters are then compared between the two subsets. By performing this split-and-compare procedure multiple times for a range of $k$, the $k$ may be optimized to reduce the discrepancy between the clusterings of the split data.
Each of these approaches was applied, with similar results, to the data presented here; we present in detail the results using the BIC-based method.

Once \( k \) and \( l \) have been assigned, clustering of the data embedded in the \( l \) eigenvectors may be clustered using \( k \)-means [4]. The spectral clustering procedure offers several advantages over simple clustering of the original data using \( k \)-means: first, the Fiedler vector provides a natural means to estimate the number of clusters; and second, because spectral clustering operates on similarity of the samples, rather than planar cuts of the high-dimensional feature space, complex correlation structures can be identified. A full discussion of its advantages is given in [13, 14].

2. Scrubbing

After the clustering step has been performed and each data point assigned to a cluster, we wish to “scrub out” the portion of the data explained by those clusters and consider the remaining variation. This is done by computing first the cluster centroids (that is, the mean of all the datapoints assigned to a given cluster), and then subtracting the data’s projection onto the centroids from the data itself, yielding the residuals. The clustering step may then be repeated on the residual data, revealing structure that may exist at multiple levels, until either a) the eigenvalues of the Laplacian in the scrubbed data are indistinguishable from a null model as described above; or b) the cluster centroids are linearly dependent. (It should be noted here that the residuals may still be computed in the latter case, but it is unclear how to interpret linearly dependent centroids.)

B. Average Commute Time

The average commute time between nodes \( i \) and \( j \), \( \eta(i,j) \), is the average number of steps a random walker, starting at node \( i \), will take to visit node \( j \), and return to node \( i \). This provides a useful measure of distance on a weighted graph. The average commute time can be used to find effective potential wells on a graph using this distance, and thus facilitates clustering of nodes.

Following the work in ref. [11] in order to calculate \( \eta(i,j) \) we first calculate the pseudoinverse of the Laplacian matrix, \( L \). We can use the Moore-Penrose pseudoinverse to calculate the pseudoinverse as

\[
L^+ = (L - ee^T/n)^{-1} + ee^T/n,
\]

where \( e \) is a column vector of ones and \( n \) is the number of nodes in the graph.

We now define the node vectors \( e_i \) to be

\[
e_i = [0, ..., 1_i, ..., 0_n]^T,
\]

and the volume of the graph \( V_G \) to be

\[
V_G = \sum_{i}^{n} \sum_{j}^{n} a_{ij},
\]
where \( a_{ij} \) are the elements of the adjacency matrix of the graph.

The commute time can now be calculated as

\[
\eta(i,j) = V_{G}(e_i - e_j)^T L^+ (e_i - e_j).
\] (6)

Using the eigenvector decomposition of \( L^+ \), the node vectors can be mapped into a Euclidean space that preserves the commute time distance, where \( x_i = U^T e_i \) and \( U \) is an orthonormal matrix composed of the eigenvectors of \( L^+ \) in order of decreasing associated eigenvalues \( \lambda_k \). These new vectors are then scaled as \( x_i' = \Lambda^{1/2} x_i \), where \( \Lambda = \text{Diag}(\lambda_k) \), and \( \eta(i,j) \) is recalculated as

\[
\eta(i,j) = V_{G}(x_i' - x_j')^T L^+ (x_i' - x_j').
\] (7)

Once this is done we can compute the optimal dimension reduction and number of clusters using the methods discussed in reference to PDM.

III. DATA

The data used in this study is derived from gene-expression profiling study of radiation toxicity designed to identify the determinants of adverse reaction to radiation therapy [19]. The gene expression data is publicly available through the Gene Expression Omnibus [20] repository.

Radiation therapy is used to treat over 60% of cancer patients, and radiation toxicity affects 5–10% of treated individuals significantly enough to warrant stopping treatment. To investigate the radiation response in sensitive and non-sensitive patients, lymphocytes from a total of 57 individuals were obtained [19]. The 57 individuals comprised four groups: 14 cancer patients with significant radiation sensitivity; 13 cancer patients with little or no radiation sensitivity; 15 healthy subjects with no history of cancer; and 15 subjects with a diagnosis of skin cancer before the age of 40. (Because skin cancer is associated with altered response to UV radiation, the latter group was included for comparison.) The cells were then subject to three treatments each: UV radiation exposure; ionizing radiation (IR) exposure; and “mock” treatment, in which the cells were placed in the same suspension as the other treatments, but not irradiated [19]. The study thus has a 4x3 design comprising 171 samples. RNA from these cells was hybridized to Affymetrix HGU95AV2 chips, providing gene expression data for each sample for 12615 unique probes. The microarray data was normalized using RMA [21]. In the original work [19], the authors identified 24 genes which were able to predict acute radiation toxicity with 64.2% sensitivity.

IV. PRELIMINARY RESULTS

A. Partition decoupling

Using spectral clustering to classify the cells described above yields precise classification of treatment groups, independent of the sensitivity group. Here, we use the BIC optimization method
to determining the number of clusters $k$ and resampling of the data to determine the dimension of the embedding $l$. 100 $k$-means runs were performed, choosing the clustering yielding the smallest within-cluster sum of squares. Classification results are given in Table 1 and Figure 1; the clustering assignments correspond exactly to the exposure categories.

In order to compare the performance of spectral clustering to that of $k$-means, we ran $k$-means on the original data using $k = 3$ and $k = 4$, corresponding to the number of treatment groups and number of cell type groups respectively. As with the spectral clustering, 100 random $k$ means starts were used, and the smallest within-cluster sum of squares was chosen. The results, given in Tables 2 and 3, show substantially noisier classification than the results obtained via spectral clustering. It should also be noted that the number of clusters $k$ used here was not derived from the characteristics of the data, but rather assigned in a supervised way that requires additional knowledge of the probable number of categories (here, dictated by the study design).

While the pure $k$-means results are noisy, the $k = 4$ classification yields a cluster that is dominated by the highly radiation-sensitive cells (cluster 4, Table 3). Membership in this cluster versus all others identifies highly radiation-sensitive cells with 62% sensitivity and 96% specificity; if we restrict the analysis to the clinically-relevant comparison between the last two cell types—that is, cells from cancer patients who show little to no radiation sensitivity and those from cancer patients who show high radiation sensitivity—the classification identifies radiation-sensitive cells with 62% sensitivity and 82% specificity. (For comparison, note also that in [19], the authors were able to obtain 64.2% sensitivity with a reduced gene set.)

These results suggest that there exist cell-type specific differences in gene expression between the high radiation sensitivity cells and the others. To investigate this, we perform the “scrubbing” step of PDM, taking only the residuals of the data after projecting onto the clusters obtained in the first pass. Since the first level of clustering corresponds precisely to treatment type, clustering on the scrubbed data should reveal cell-type specific differences that are independent of the treatment. Once again, we use the BIC optimization method to determine the number of clusters $k$ and resampling of the data to determine the dimension of the embedding $l$. This time, two clusters are found to be optimal; classification results are given in Table 4 and Figure 2. As in the $k$ means, one cluster is dominated by radiation-sensitive cells, but the classification sensitivity is much higher (83%) without a large sacrifice in specificity (92% for all samples, 72% when comparing solely to low radiation-sensitivity patients). This sensitivity is considerably greater than the 62.4% obtained in the initial analysis [19], suggesting that there exist patterns of gene expression that are able to distinguish the radiation-sensitive patients which were not identified in [19].

Also as in the pure $k$-means results, no distinction is seen between the healthy skin fibroblasts and those of skin cancer patients, who were expected to show altered UV response; patients who had little to no radiation sensitivity like between the (insensitive) healthy and skin-cancer-positive control groups and the highly radiation-sensitive groups. Unfortunately, because more finely detailed data on the radiation sensitivity of the subgroups is not available, it is not possible here to state whether the individuals in the low sensitivity group who were clustered with the high sensi-
tivity group had higher radiation sensitivity than those who did not. Further scrubbing resulted in residuals that were indistinguishable from noise (see Methods) and we conclude that only two levels of structure – corresponding to exposure and cell type – are resent in the data.

B. CTD results

While spectral clustering of the skin cells allows all three of the treatments to be perfectly determined, an interesting result is obtained by attempting to cluster the data with two clusters. In Fig. 3 it can be seen that the clustering separates the UV treatment into one cluster, and the IR and Mock treatments into a separate cluster. Using $k$-means these groups are again perfectly divided with all the UV treatment cases in one cluster, and the other treatments in a separate cluster.

Using $k$-means again, though this time with the CTD yields a different result however. The clustering now divides the IR treatment into one cluster, and the UV and mock treatments into a separate cluster. This divide is not quite as clean as for the spectral clustering though, with approximately 3.5% of the UV treatment points being placed into the IR cluster, and 3.5% of the IR treatment points being placed in the UV/mock cluster, Fig. 4. This could mean that there are many more shared paths between the UV and mock treatments than by the IR and mock treatments. It could also mean that there are a large number of paths between IR treatment points, resulting in a much tighter clustering of that portion of the data.

V. DISCUSSION AND FUTURE WORK

Our preliminary analysis of gene microarrays using PDM and CTD techniques have shown promising results. Neither of these techniques relies on heuristics or a curated set of genes; rather, the data itself is used to determine both the number of clusters and the dimensionality to which the $10^5$-probe data set is reduced. They therefore provide an unsupervised method for discovering communities in genes and people. In at least one case they have correctly partitioned three different treatments into separate groups.

The results of PDM applied to the data permit us to conclude that two layers of structure, corresponding to radiation exposure and cell type, are present in the gene expression data. While cell-type is weakly discernable using traditional clustering methods on the original data, we find that scrubbing out the exposure-related structure reveals a much cleaner clustering of cell-type using spectral clustering. Notably, PDM not only identifies exposure groups with 100% accuracy, but also permits us to improve considerably the classification of radiation-sensitive cells to 83% from the 64% sensitivity reported in [19]. This suggests that there exist strong patterns, previously undetected, of gene expression that correlate with radiation exposure and cell type.

The different measures of distance used in this paper, using the Laplacian vs. using the CTD, have been shown to result in different clustering choices. This points to particular features of
the space we are working in. There are a number of different mechanisms that could cause this however.

In the case of the skin data we would like to find the reason that using the Laplacian results in the UV and Mock treatments being clustered together, while the CTD clusters the IR and Mock treatments together. This could result from a number of different correlations between the treatments. This is due to the fact that when using the CTD adding more paths between nodes, as well as strengthening the links between nodes will shorten the distance between them, as measured by the CTD. The Laplacian does not have this first property. This points to the different clustering choices between these methods indicating something about the topology of the underlying network. More work is required however to discover what this actually means, and will be a focus of future research.

Several avenues for further investigation exist. First, the analyses performed here rely on the computation of the Pearson correlation coefficient as a measure of pairwise similarity between samples. Because it is well-established that the Pearson correlation is sensitive to outliers, use of the Spearman rank correlation coefficient or mutual information may be more appropriate; the sensitivity of the clustering to the pairwise-similarity metric needs to be fully explored. Second, we presented several approaches for optimizing the number of clusters and the embedding dimension, and a comparison of these schemes would aid future investigators.

Acknowledgements

The authors would like to thank Greg Leibon, Scott Pauls, and Dan Rockmore for their generous, extensive advice and assistance. RB would further like to thank Sean Brocklebank for many fruitful discussions. This work was made possible by the Santa Fe Institute Complex Systems Summer School (2009). RB is supported through the Cancer Prevention Fellowship Program and a Cancer Research Training Award, National Cancer Institute, NIH.


### Tables and Figures

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mock</strong></td>
<td>57</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td>IR</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td><strong>UV</strong></td>
<td>0</td>
<td>0</td>
<td>57</td>
</tr>
</tbody>
</table>

**TABLE 1:** Spectral clustering of expression data versus exposure; exposure categories are reproduced exactly.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mock</strong></td>
<td>36</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td>IR</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td><strong>UV</strong></td>
<td>3</td>
<td>14</td>
<td>40</td>
</tr>
</tbody>
</table>

**TABLE 2:** $k$=means clustering of expression data versus exposure using $k = 3$.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>19 18 8 0</td>
</tr>
<tr>
<td>Skin cancer</td>
<td>8 23 14 0</td>
</tr>
<tr>
<td>Low radiation sensitivity</td>
<td>13 11 8 7</td>
</tr>
<tr>
<td>High radiation sensitivity</td>
<td>6 1 9 26</td>
</tr>
</tbody>
</table>

**TABLE 3:** $k$=means clustering of expression data versus cell type using $k = 4$.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>45 0</td>
</tr>
<tr>
<td>Skin cancer</td>
<td>45 0</td>
</tr>
<tr>
<td>Low radiation sensitivity</td>
<td>28 11</td>
</tr>
<tr>
<td>High radiation sensitivity</td>
<td>7 35</td>
</tr>
</tbody>
</table>

**TABLE 4:** Spectral clustering of exposure data with exposure-correlated clusters scrubbed out, versus cell type.
FIG. 1: Laplacian matrix eigenvalues (top) and Fiedler vector values (bottom) for spectral clustering of radiation exposure data. In the top plot, the resampling-based threshold for eigenvalue significance is shown in cyan, with smaller eigenvalues plotted in red. In the bottom plot, we show each sample's Fiedler vector value along with the resulting clustering. A Gaussian mixture fit to the density (bottom right) of the Fieldler vector indicates three clusters; the resulting cluster assignment for each sample is indicated by color. True treatment categories of each sample are given as shapes: crosses denote mock; circles, UV; triangles, IR. The four cell types (healthy, skin cancer, radiation insensitive, radiation sensitive) are separated by vertical lines. It can be seen that the cluster assignment correlates precisely with the exposure type, independent of cell type.
FIG. 2: Laplacian matrix eigenvalues (top) and Fiedler vector values (bottom) for spectral clustering of scrubbed radiation exposure data. In the top plot, the resampling-based threshold for eigenvalue significance is shown in cyan, with smaller eigenvalues plotted in red. In the bottom plot, we show each sample’s Fiedler vector value along with the resulting clustering. A Gaussian mixture fit to the density (bottom right) of the Fiedler vector indicates two clusters; the resulting cluster assignment for each sample is indicated by color. True treatment categories of each sample are given as shapes: crosses denote mock; circles, UV; triangles, IR. The four cell types (healthy, skin cancer, radiation insensitive, radiation sensitive) are separated by vertical lines. It can be seen that the cluster assignment correlates loosely with the final (radiation sensitive) cell type.
FIG. 3: Two-dimensional projection of gene expression data, clustered using $k$-means ($k = 2$) on the Laplacian. Open circles correspond to treatment type, and the symbols within the circles designate which cluster they belong to. It can be seen that the IR and Mock treatments lie in the same cluster, while the UV treatment lies in a separate cluster.
FIG. 4: Two-dimensional projection of gene expression data, clustered in two groups using the CTD method. Open circles correspond to treatment type, and the symbols within the circles designate which cluster they belong to. Contrary to the clustering using the Laplacian, the UV and Mock treatments lie in the same cluster, while the IR treatment lies in a separate cluster. In this case however 2 of 57 UV treatment points lie in the IR cluster, and 2 of 57 IR treatment points lie in the UV cluster.