BIOLOGICAL PATHWAY INFERENCE USING MANIFOLD EMBEDDING

Arvind Rao

Lane Center for Computational Biology
Carnegie Mellon University,
Pittsburgh, PA 15213, USA
*email: ukarvind@cmu.edu

Alfred O. Hero, III

EECS and Bioinformatics,
University of Michigan,
Ann Arbor, MI 48109, USA,
email: hero@umich.edu

ABSTRACT

Disease occurs due to aberrant modulation of biological pathways. Identification of activated gene pathways from gene expression data is an important problem. In this work, we develop a framework identifying activated pathways that incorporates cellular location of the gene, using gene ontology databases, in addition to gene expression data. This information is combined using Laplacian Eigenmaps to co-embed these data into a low dimensional manifold. Model-based clustering is then performed to identify biologically relevant activated pathways in the gene expression data. We illustrate the effectiveness of our manifold embedding approach for the problem of extracting immune system pathways from a macrophage gene expression dataset [11].

Index Terms—functional data analysis (FDA), gene ontology (GO), immune response, Laplacian eigenmaps, heterogeneous data integration.

I. INTRODUCTION

Biological processes are orchestrated by a precise spatio-temporal expression of genes. Systemic disease can occur due to mis-expression of several genes (i.e. gene sets or pathways). Identification of these sets of aberrant pathways is an important problem as it can lead to diagnostic and therapeutic procedures. There is an ongoing effort to find inhibitors that can target disease-implicated pathways. For example, two well-known drugs Gleevec and Tarceva target receptor tyrosine kinase signaling pathways that have a reported role in cancer [6].

In this work, we present a method to identify pathways that are modulated in disease progression. Using time series expression profiles of gene expression, we use functional data analysis (FDA) [10] to process, analyze and cluster the data into pathways. Instead of trying to identify modulated pathways by directly clustering gene expression data, we first embed the gene expression into a manifold that imposes certain biological constraints, specifically, the physical proximity of pairs of genes within the cell as determined by publicly available gene ontology information. The obtained embedding produces a clustering that is much more concordant with true pathway membership. This methodology is a principled way to incorporate prior biological knowledge into bioinformatics analysis. We use gene expression and gene ontology for identifying immune response pathways in a macrophage gene expression dataset [11] to illustrate our approach.

II. DATA EXTRACTION AND PRE-PROCESSING

One of the most common processes involving gene-pathway modulation is the systemic response of the immune system to an invading pathogen. For this study, we use gene expression data gathered by the Young group [11]: this data profiles the various gene activation programs initiated in macrophage cells on exposure to various pathogens such as tuberculosis, E. coli and Staphylococcus aureus. As a control, latex beads coated without bacteria are presented to the macrophage cell population. In this work, we study the differential pathway activity between tuberculosis and control conditions. Our methods are of course more general and can be applied to other conditions as well.

The dataset contains expression values at 8 time points for 168 unique macrophage genes. These correspond to gene expression profiling at times 0.5, 1, 3, 6, 12, 16, 18, and 24 hours after exposure to the pathogen (or control). Since macrophages exhibit an early innate as well as late adaptive immune response, it is of interest to identify which genes are expressed early or late in the host response program.

Of the many possible interpolation schemes used in functional data analysis [10], we used B-spline basis functions. Under this representation, we have,

\[ x_i(t_j) = \sum_{k=1}^{K} c_k B_k(t_j) \] for \( i = 1, 2, \ldots, n \) with \( n = 168 \), and \( B_k(t_j) \) are B-spline basis functions of order 3.

III. CLUSTERING WITH FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS (FPCA)

A popular method to identify pathways in functional data is to cluster the temporal profiles in a low dimensional space using a linear embedding like PCA ([7], [14]). However, as shown in this section, linear embeddings inadequately
separate gene-expression clusters and, as shown in the next section, more sophisticated non-linear embeddings that account for gene ontology are significantly better. Specifically, functional PCA (fPCA) approach projects this functional data into two-dimension to enhance interpretability. Functional PCA aims to find a solution to the eigenvalue problem \[ \phi \lambda \mathbf{b} = \lambda \mathbf{b} \], where \( C = \sum_{i=1}^{n} \phi (\mathbf{x}_i, \mathbf{x}_i) / n \), \( \phi = [(B_k, B_m)] \) and \( \mathbf{b} = (b_1, b_2, \ldots, b_k) \). The \( j \)th principal component eigenvector \( b_j \) of \( \phi \) leads to an estimate \( \epsilon_j = [B_1, B_2, \ldots, B_K]^{T} \mathbf{b}_j \) of the eigenfunction. With this, the \( j \)th principal component score is given by \( \alpha_{i,j} = (\mathbf{x}_i, \epsilon_j) \).

The set of scores \( \{\alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,p}\} \in \mathbb{R}^p; i = 1, 2, \ldots, 168 \) is clustered to identify disparate groups of similar gene expression profiles [14]. The Mixture-of-Gaussian (MoG) framework [8], Bayesian Information criterion (BIC) is used to determine the optimal number of clusters. The result of applying two dimensional \( (p = 2) \) fPCA followed by MoG clustering, on the tuberculosis data is illustrated in Fig. 1 for the Young data set [11].

![Fig. 1. BIC based model clustering (four clusters) for fPCA embedded tuberculosis data (without semantic similarity).](image)

The deficiency of the PCA linear embedding is revealed in the poor match of the MoG ellipsoidal assumption to the data; specifically the largest ellipse in Fig. 1. This poor clustering performance can be quantified by examining cluster purity, defined as the number of co-clustered genes that are in the same cellular location. This is done using the FATIGO+ tool at http://babelomics.bioinfo.cipf.es/fatigoplus/. The results of this analysis is shown in Fig. 2. This figure indicate that only about 50% of the co-clustered genes are co-located (in the nucleus). Biologically, the cellular proximity of two genes is essential for their interaction along a pathway. Thus, unless cellular proximity can be explicitly incorporated into the clustering framework, clustering can potentially fail in the discovery of true pathway components ([7],[1]). Location proximity can be obtained by generating the GO semantic-similarity matrix [4] between the 168 genes.

![Cellular component. Level: 0](image)

**IV. CLUSTERING WITH GO SIMILARITY**

**GO semantic similarity:** The semantic similarity of any two GO terms in the ontology hierarchy is based on the number of shared parents and the information content of the individual GO terms. We use the Jiang-Conrath similarity measure [3], given by,

\[
W_{i,j} = \text{sim}(c_i, c_j) = \frac{1}{J_{\text{dist}}(c_i, c_j)}, \quad \text{with} \quad J_{\text{dist}}(c_i, c_j) = 2\log(p(\text{iso}(c_i, c_j))) - \left[\log(p(c_i)) + \log(p(c_j))\right],
\]

where \( c_i \) and \( c_j \) are two terms (nodes) in the GO ontology tree \( (i, j \in \{1, 2, \ldots, 168\}) \). \( \text{iso}(c_i, c_j) \) refers to the the information content of the last common parent of these two nodes. The information content is computed based on the probabilities of observing the individual nodes and their last common ancestor in an overall corpus [4].

The next step involves the use of manifold embedding techniques that can integrate GO similarity with expression-level similarity to construct a co-embedding of the genes as points in 2D space. One such technique is to use Laplacian Eigenmaps [2], that can capture both these relationships (semantic and expression similarities).

**LLE (Laplacian Eigenmaps) co-embedding procedure**

- Build the \( K \times K \) \( (K = 168) \) dimensional weight matrix \( W \) from the Gene Ontology (“Cellular Component”) terms of the genes in the dataset. This distance is the “normalized” semantic similarity alluded to above.
- Assign weight \( W_{i,j} \), from (1) for each gene pair \( (i,j) \), for each of the \( \binom{K}{2} \) gene pairs. Note: The higher this weight, the closer the genes are.
- Find \( n \) nearest neighbors using the euclidian distance in fPCA space. The scores of the functional data along the first two principal components can be interpreted as co-ordinates in a euclidian space.

![Fig. 2. GO purity of clusters from fPCA.](image)
• Form the Graph Laplacian:
\[
L_{i,j} = \begin{cases} 
  d_i = \sum_k W_{i,k} & \text{if } i = j; \\
  -W_{i,j} & \text{if } i \text{ is connected to } j; \\
  0 & \text{otherwise}. 
\end{cases}
\]

• Solve: \( \min_y y^T L y = \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 W_{i,j} \) (2), subject to:
  - \( y^T D y = 1 \), and
  - \( y^T D_1 y = 0 \),

where \( D_{i,i} = \sum_j W_{i,j} \), a diagonal weight matrix.

• Embed the co-ordinates to a lower dimensional manifold, using the solution (the Laplacian Eigenmap) obtained from the minimization above.

  - The solution to (2) is given by the \( d \) generalized eigenvectors associated with the \( d \) smallest generalized eigenvalues solving \( L y = \lambda D y \) (neglecting the zero eigenvalue and its eigenvector).
  - If \( y = [y_1, \ldots, y_d] \) is the collection of these eigenvectors, then the embedding is given by: \( y_i = (y_1, \ldots, y_d)^T \), i.e., the \( d \) dimensional representation of the \( i^{th} \) data point (gene).

• For visualization purposes, we take dimensionality \( d = 2 \). The number of neighbors was set at \( n = 5 \). The final embedding of the functional data based on expression and location modalities is shown in Fig. 3(a).

V. RESULTS:

After LLE embedding the original scores (from fPCA) based on the GO semantic similarity matrix \( W \), and applying MoG clustering we obtain the embedding shown in Fig.3(a). Note that this is a very different embedding of the data compared to Fig.1 MoG. An examination of the GO ("cellular component" annotation) is shown in Fig.3(b). This indicates that the GO enrichment of genes that are co-located in the cell (like in nucleus) is much higher (~70% compared to 55% before).

In addition to this improved biological enrichment, we compare the results of clustering with and without co-embedding by examining three measures of clustering quality— the Dunn index, the Davies-Bouldin index and the connectivity index.

1) The Dunn’s validity index \( D \) is defined as \( D = \min_{1 \leq i < n} \frac{\min_{1 \leq k \leq n \neq i} \{ \frac{d(x_i, x_k)}{\max(d(x_i, x_k))} \}}{\min_{1 \leq k \neq j} \{ \frac{d(x_i, x_j)}{\max(d(x_i, x_j))} \}} \), with the numerator denoting the intercluster distance and the denominator being the within-cluster distance. A higher value is better.

2) The Davies-Bouldin index is a ratio of the sum of within-cluster-scatter to between-cluster scatter. This should have a small value for good clustering.

3) If \( n_{i,j} \) is denoted as the \( j^{th} \) nearest neighbor of observation \( i \), and \( x_{i,n_{i,j}} \) is 1/j if \( i \) and \( n_{i,j} \) are in different clusters, and 0 otherwise – the connectivity index for a particular clustering partition is defined as, \( Conn = \sum_{i=1}^{N} \sum_{j=1}^{K} x_{i,n_{i,j}} \), where \( N \) is the number of observations and \( K \) is the number of clusters in the partition. A low value indicated well-separated clusters.

Each of these metrics quantifies the internal stability of the clusters obtained. As shown in Table. I, we obtain superior values of all these stability measures in the co-embedded clustering (shown in Fig. 3a) as compared to the standard linear embedding in Fig 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dunn index</th>
<th>DB index</th>
<th>connectivity index</th>
</tr>
</thead>
<tbody>
<tr>
<td>without co-embedding</td>
<td>0.59</td>
<td>0.65</td>
<td>9.83</td>
</tr>
<tr>
<td>with co-embedding</td>
<td>3.22</td>
<td>0.44</td>
<td>2</td>
</tr>
</tbody>
</table>

Table I. Dunn, Davies-Bouldin and connectivity indices for clustering with (Fig. 1) and without co-embedding (Fig. 3a).

Finally, we examine results of cluster concordance between true pathway memberships and cluster-memberships with and without co-embedding. We use several standard metrics from literature [13]. Although we obtained these metrics for all pathways in the KEGG database [9], in the Tables. II and III – we only show those pathways that have at least 10 genes in common with the immune-response dataset [11]. As shown, several important pathways important to pathogenesis show improved clustering after co-embedding.

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Number of genes</th>
<th>Rand Index</th>
<th>Folkkes-Mallow Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xenobiotic degradation</td>
<td>20</td>
<td>0.226</td>
<td>0.475</td>
</tr>
<tr>
<td>Cytokine-Cytokine interaction</td>
<td>24</td>
<td>0.202</td>
<td>0.450</td>
</tr>
<tr>
<td>Chemokine signaling</td>
<td>13</td>
<td>0.141</td>
<td>0.373</td>
</tr>
<tr>
<td>Toll-like Receptor</td>
<td>12</td>
<td>0.166</td>
<td>0.408</td>
</tr>
<tr>
<td>NOD receptor</td>
<td>14</td>
<td>0.241</td>
<td>0.491</td>
</tr>
<tr>
<td>Hematopoietic (B-cell/T-cell)</td>
<td>10</td>
<td>0.177</td>
<td>0.421</td>
</tr>
<tr>
<td>Cytokine-signaling</td>
<td>16</td>
<td>0.325</td>
<td>0.570</td>
</tr>
</tbody>
</table>

Table II. Rand and Folkes-Mallow indices comparing true gene-memberships with clustering without co-embedding

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Number of genes</th>
<th>Rand Index</th>
<th>Folkkes-Mallow Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xenobiotic degradation</td>
<td>20</td>
<td>0.373</td>
<td>0.611</td>
</tr>
<tr>
<td>Cytokine-Cytokine interaction</td>
<td>24</td>
<td>0.264</td>
<td>0.514</td>
</tr>
<tr>
<td>Chemokine signaling</td>
<td>13</td>
<td>0.243</td>
<td>0.493</td>
</tr>
<tr>
<td>Toll-like Receptor</td>
<td>12</td>
<td>0.212</td>
<td>0.460</td>
</tr>
<tr>
<td>NOD receptor</td>
<td>14</td>
<td>0.329</td>
<td>0.574</td>
</tr>
<tr>
<td>Hematopoietic (B-cell/T-cell)</td>
<td>10</td>
<td>0.288</td>
<td>0.537</td>
</tr>
<tr>
<td>Cytokine-signaling</td>
<td>16</td>
<td>0.341</td>
<td>0.584</td>
</tr>
</tbody>
</table>

Table III. Rand and Folkes-Mallow indices comparing true gene-memberships with clustering via co-embedding
From Tables II and III we see that the clustered non-linear co-embedding which respects cellular location as well as expression better identifies closely interacting pathway components. This suggests that our proposed non-linear co-embedding approach can be used to generate more biologically relevant hypotheses.

VI. CONCLUSIONS

In this work, we have demonstrated a method to infer pathway components. Using Laplacian Eigenmaps, we co-embed genes based on gene expression and cellular location. Model based clustering, and several metrics of cluster concordance confirm that co-embedded data result in higher cluster purity with respect to cellular location.

The overall contribution of this work is an approach that combines biological prior knowledge to produce a manifold where the correlation in expression is concordant with true pathway membership. This approach combines functional data analysis on expression data with ontology to yield biologically relevant results via heterogeneous data integration. Though there has been some previous work [1] combining gene expression with ontology to understand gene co-regulation, we are aiming to do this for entire pathways or gene sets.

VII. ACKNOWLEDGEMENTS

A.R was supported by the Lane Fellows program at Carnegie Mellon University. This work was partially supported by a grant from the National Science Foundation, grant number CCF 0830490.

VIII. REFERENCES